



UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

SESSION 2020/2021 SEMESTER 2

SEC12143 – PROBABILITY & STATISTICAL DATA ANALYSIS

Dr. Nor Azizah Ali

Project 2

SECTION: 05

Group Name: Data Miners

List of Members:

Name	Matric No.
1. Omar Mokhtar Bin Yohan	A20EC0128
2. Muhammad Sulaiman Daud Syu'aib Bin Yaacob	A20EC0100
3. Ibtesham Ahmed Promit	A20EC4027
4. Md Mirza Shihab	A16CS4016

Introduction

In this day and age, we spend a significant amount of our time on the road travelling from one place to another. It is obvious that transportation enhances people's living standards while also contributing to a country's economic prosperity. It makes it easier for people to move themselves and their belongings across the world to complete their jobs and also move goods between places to where it is demanded. In this study, our group has conducted several analyses and tests to identify relationships between accidents and factors. The analyses used in this study are hypothesis testing on 2 sample test, correlation test, regression test, goodness of fit test, and ANOVA test. The hypothesis testing on 2 sample test is a hypothesis test that can assist in determining if a variation in estimated proportions reflects a difference in population proportions. The difference between two proportions follows a roughly normal distribution. In general, the null hypothesis asserts that the two proportions are equal. Correlation is a bivariate study that determines the intensity of connection and the direction of the link between two variables. The correlation coefficient has a value between +1 and -1 depending on the strength of the link. A value of 1 shows that the two variables are perfectly linked. Regression analysis is a collection of statistical methods for estimating connections between one or more independent variables and a dependent variable. It may be used to analyse the strength of the link between variables and to forecast their future relationship. There are numerous types of regression analysis, including linear, multiple linear, and nonlinear. Simple linear and multiple linear models are the most popular. Nonlinear regression analysis is often employed for more sophisticated data sets with a nonlinear connection between the dependent and independent variables. The goodness-of-fit test is a statistical hypothesis test that determines how well sample data match a normal distribution from a population. To put it another way, this test determines if your sample data is representative of the data you would expect to find in the actual population or whether it is biased in any way. The goodness-of-fit test determines the difference between the actual values and those that would be anticipated of the model in the case of a normal distribution. The one-way ANOVA analyses the means of the groups in question and evaluates if any of them are statistically substantially different from one another. It specifically examines the null hypothesis.

Dataset

The dataset that we chose is about bad drivers in various states. In the dataset we can see the names of 51 states on the first column. The other seven columns represents Number of drivers involved in fatal collisions per billion miles, who were Speeding, who Were Alcohol-Impaired, who Were Not Distracted and those who Had Not Been Involved In Any Previous Accidents accordingly from second column to 4th column. Whereas, the last two column are about car insurance and losses incurred by insurance companies for collisions per insured driver. We did statistical analysis to find out which states have the worst drivers like hypothesis 2 sample test, Correlation test, Regression test, Goodness of fit test, Chi-Square test of independence, ANOVA test.

Data Analysis

Hypothesis Testing on 2 Sample Test

Based on the data obtained, there are two traits of drivers who are contribute to the fatal collision in 16 sample which are drivers who were not distracted while driving and the drivers who had not been involved in any previous accidents. We claim that the average of drivers who were not distracted while driving and the drivers who had not been involved in any previous accidents are equal. The hypothesis testing on mean with unknown variance was conducted. The steps are shown below.

Using **t-test** since $n \leq 30$ and case 1, **assumed equal variance** ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

Significance level: $\alpha = 0.05$

H₀: $\mu_1 = \mu_2$ (There is no difference mean between drivers who were not distracted while driving and the drivers who had not been involved in any previous accidents)

H₁: $\mu_1 \neq \mu_2$ (There is a difference mean between drivers who were not distracted while driving and the drivers who had not been involved in any previous accidents)

States	Percentage Of Drivers Involved In Fatal Collisions Who Were Not Distracted(x_1)	$(x_1 - \bar{x}_1)^2$	Percentage Of Drivers Involved In Fatal Collisions Who Had Not Been Involved In Any Previous Accidents(x_2)	$(x_2 - \bar{x}_2)^2$
Alabama	96	30.25	80	156.25
Alaska	90	0.25	94	2.25
Arizona	84	42.25	96	12.25
Arkansas	94	12.25	95	6.25
California	91	0.25	89	12.25
Colorado	79	132.25	95	6.25
Connecticut	87	12.25	82	110.25
Delaware	87	12.25	99	42.25
District of Columbia	100	90.25	100	56.25
Florida	92	2.25	94	2.25
Georgia	95	20.25	93	0.25

Hawaii	82	72.25	87	30.25
Idaho	85	30.25	98	30.25
Illinois	94	12.25	96	12.25
Indiana	95	20.25	95	6.25
Iowa	97	42.25	87	30.25
N=16	$\bar{x}_1=1448/16=90.5$	$\sum(x_1-\bar{x}_1)^2=532$	$\bar{x}_2=1480/16=92.5$	$\sum(x_2-\bar{x}_2)^2=516$

$$n = 16$$

x_1 = Percentage of drivers involved in fatal collisions who were not distracted

x_2 = Percentage of drivers involved in fatal collisions who had not been involved in any previous accidents

$$\text{Standard deviation, } s = \sqrt{\left\{ \frac{\sum(x-\bar{x})^2}{n-1} \right\}}$$

$$\text{So } S_1 = \sqrt{(532/16-1)} = 5.96$$

$$S_2 = \sqrt{(516/16-1)} = 5.87$$

Taking $\alpha=0.05$

The Test Statistics,

$$T_0 = \frac{X_1 + X_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{90.5 + 92.5 - 0}{\sqrt{\frac{5.96^2}{16} + \frac{5.87^2}{16}}} = -0.956$$

Degrees of Freedom,

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2 - 1}}$$

$$= \frac{19.1289}{0.6378}$$

$$= 29.99 \approx 30$$

Therefore, using $\alpha=0.05$, we reject H_0 if

$$t_0^* > t_{0.025,30} = 2.042 \text{ or } t_0^* < -t_{0.025,30} = -2.042$$

Conclusion:

Since $2.042 > t_0^* = -0.956 < -2.042$, thus we fail to reject the null hypothesis. That is, at 0.05 level of significance, we do not have a strong evidence to conclude that the mean percentage of drivers involved in fatal collisions who were not distracted is different from the mean percentage of drivers involved in fatal collisions who had not been involved in any previous accidents.

Correlation Test

State	Number of drivers involved in fatal collisions per billion miles (x)	Losses incurred by insurance companies for collisions per insured driver (\$), (y)	xy	x ²	y ²
Alabama	18.8	145.08	2727.5	353.44	21048.21
Alaska	18.1	133.93	2424.13	327.61	17937.24
Arizona	18.6	110.35	2052.51	345.96	12177.12
Arkansas	22.4	142.39	3189.54	501.76	20274.91
California	12	165.63	1987.56	144	27433.30
Colorado	13.6	139.91	1902.78	184.96	19574.81
Connecticut	10.8	167.02	1803.82	116.64	27895.68
Delaware	16.2	151.48	2453.98	262.44	22946.19
District of Columbia	5.9	136.05	802.695	34.81	18509.60
Florida	17.9	144.18	2580.82	320.41	20787.87
Georgia	15.6	142.8	2227.68	243.36	20391.84
Hawaii	17.5	120.92	2116.1	306.25	14621.65
Idaho	15.3	82.75	1266.08	234.09	6847.56
Illinois	12.8	139.15	1781.12	163.84	19362.72
Indiana	14.5	108.92	1579.34	210.25	11863.57
	$\sum x = 230$	$\sum y = 2030.56$	$\sum xy = 30895.66$	$\sum x^2 = 3749.82$	$\sum y^2 = 281672.27$

Value of the independent variable (x) = Number of drivers involved in fatal collisions per billion miles

Value of the dependent variable (y) = Losses incurred by insurance companies for collisions per insured driver (\$)

Pearson's product-moment correlation coefficient:

$$r = \frac{\sum xy - (\sum x \sum y) / n}{\sqrt{[(\sum x^2) - (\sum x)^2 / n][(\sum y^2) - (\sum y)^2 / n]}}$$

$$r = \frac{30895.66 - ((230)(2030.56)) \div 15}{\sqrt{\left[3749.82 - \frac{230^2}{15}\right] \left[281672.27 - \frac{2030.56^2}{15}\right]}}$$

$$r = -0.1946$$

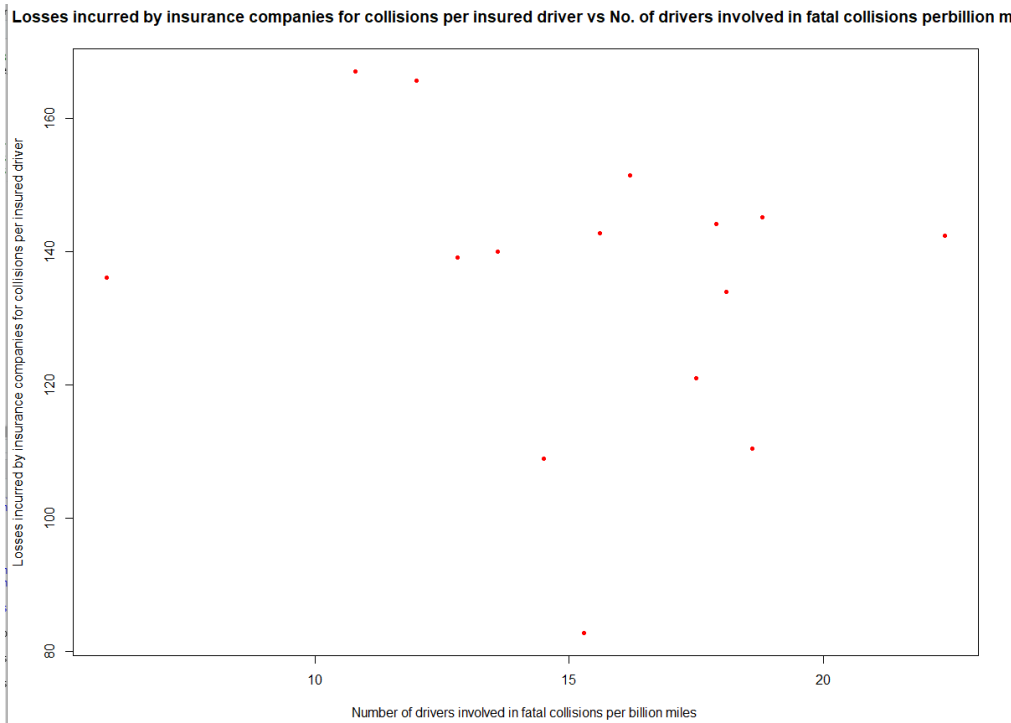


Figure 1 Scatter plot of Losses incurred by insurance companies for collisions per insured driver vs No. of drivers involved in fatal collisions per billion miles.

This correlation analysis analysed the relationship between the losses incurred by insurance companies for collisions per insured driver and the number of drivers involved in fatal collisions per billion miles. Since the dataset was in the ratio form, Pearson's technique was applied in calculating the Person's product-moment correlation coefficient. By using the R studio to calculate the coefficient correlation, r to measure the statistical relationship between these two variables. Assuming normality in both variables. We found that the sample correlation coefficient, r generated by the RStudio is -0.1946 . Since r is a negative value close to 0, there is a negative relationship between the number of drivers involved in fatal collisions per billion miles and the losses incurred by insurance companies for collisions per insured driver and there is a weak correlation between the two variables.

Significance Test for Correlation

Then, these two variables are used to test whether there is any evidence of linear relationship between them at 0.05 level of significance.

H_0 is assumed that the losses incurred by insurance companies for collisions per insured driver and the number of drivers involved in fatal collisions per billion miles has no linear correlation.

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

Where, ρ (rho) = population correlation coefficient

The test statistics formula is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, where sample size, $n = 15$ and correlation coefficient, $r = -0.1946$. The RStudio generated the value where $t = -0.71531$, $df = 13$ and $p\text{-value} = 0.4871$. At 95% confidence interval, the correlation coefficient, r is in between -0.6427853 and 0.3528366 . Since, the $p\text{-value}$ is greater than the alpha value ($0.4871 > 0.05$). So, fail to reject null hypothesis. There is a strong evidence to conclude that there is no linear correlation between the losses incurred by insurance companies for collisions per insured driver and the number of drivers involved in fatal collisions per billion miles.

Regression Test

State	Number of drivers involved in fatal collisions per billion miles (x)	Losses incurred by insurance companies for collisions per insured driver (\$), (y)	xy	x ²
Alabama	18.8	145.08	2727.5	353.44
Alaska	18.1	133.93	2424.13	327.61
Arizona	18.6	110.35	2052.51	345.96
Arkansas	22.4	142.39	3189.54	501.76
California	12	165.63	1987.56	144
Colorado	13.6	139.91	1902.78	184.96
Connecticut	10.8	167.02	1803.82	116.64
Delaware	16.2	151.48	2453.98	262.44
District of Columbia	5.9	136.05	802.695	34.81
Florida	17.9	144.18	2580.82	320.41
Georgia	15.6	142.8	2227.68	243.36
Hawaii	17.5	120.92	2116.1	306.25
Idaho	15.3	82.75	1266.08	234.09
Illinois	12.8	139.15	1781.12	163.84
Indiana	14.5	108.92	1579.34	210.25
Iowa	15.7	114.47	1797.18	246.49
Kansas	17.8	133.8	2381.64	316.84
Kentucky	21.4	137.13	2934.58	457.96
Louisiana	20.5	194.78	3992.99	420.25
Maine	15.1	96.57	1458.21	228.01

Dependent variable, (y) = Losses incurred by insurance companies for collisions per insured driver (\$).

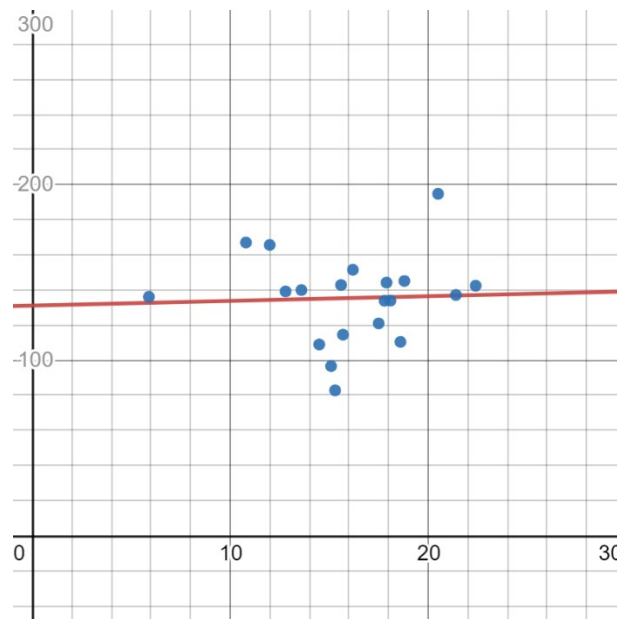
Independent variable, (x) = Number of drivers involved in fatal collisions per billion miles

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{43460.24 - \frac{(320.5)(2707.31)}{20}}{5419.37 - \frac{(320.5)^2}{20}} = \frac{75.5973}{283.358} = 0.2667$$

$$b_0 = 131.0915$$

Therefore, the regression equation is: $\hat{y} = 131.0915 + 0.2667x$

Losses incurred by insurance companies for collisions per insured driver (\$) vs Number of drivers involved in fatal collisions per billion miles



$$R^2 = \frac{SSR}{SST} = 0.001643$$

Meaning, 0.16% of the variation in Losses incurred by insurance companies for collision per insured driver explained by variation in Number of drivers involved in fatal collisions per billion miles. Thus, it has a very weak relationship or almost no relationship.

$$S_e = 26.09, S_{b1} = 1.5498$$

These indicates that the variation of observed y values is slightly inaccurate and less steep because the sample standard error of the estimate, S_e is large and the estimate of the standard error of the least squares slope is small.

Goodness Of Fit Test

State	Percentage Of Drivers Involved In Fatal Collisions Who Were Not Distracted (O)	Expected Value (E)	$\frac{(O - E)^2}{E}$
South Carolina	96	88.47	0.6409
North Dakota	99	88.47	1.2533
West Virginia	97	88.47	0.8224
Arkansas	94	88.47	0.3457
Montana	84	88.47	0.2258
Kentucky	78	88.47	1.2391
Louisiana	73	88.47	2.7051
Oklahoma	92	88.47	0.1408
Tennessee	82	88.47	0.4732
Texas	91	88.47	0.0724
South Dakota	98	88.47	1.0266
Alabama	96	88.47	0.6409
Arizona	84	88.47	0.2258
New Mexico	67	88.47	5.2104
Pennsylvania	96	88.47	0.6409
Total	1327		15.6633

$$n = 1327$$

The the number of different categories or outcomes is, $k = 15$.

Observed value (O) = Percentage of drivers involved in fatal collisions who were not distracted.

$$\text{Expected value (E)} = E = \frac{n}{k} = \frac{1327}{15} = 88.47$$

$$H_0 : P_1 = P_2 = P_3 = P_4 = P_5 = P_6 = P_7 = P_8 = P_9 = P_{10} = P_{11} = P_{12} = P_{13} = P_{14} = P_{15}$$

H_1 : At least 1 of the 15 states have percentages is different from others.

The test statistic value is :

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 15.6633$$

The significance level is :

$$\alpha = 0.05$$

Critical Value :

$$\chi^2_{14,0.05} = 23.685$$

Conclusion: Since $\chi^2 = 15.6633 < \chi^2_{14,0.05} = 23.685$, we fail to reject H_0 . Therefore we accept the claim that all of the 15 states have equal percentages.

ANOVA Test

State	Percentage Of Drivers Involved In Fatal Collisions Who Were Not Distracted (x ₁)	Percentage Of Drivers Involved In Fatal Collisions Who Had Not Been Involved In Any Previous Accidents(x ₂)
South Carolina	96	81
North Dakota	99	86
West Virginia	97	87
Arkansas	94	95
Montana	84	85
Kentucky	78	76
Louisiana	73	98
Oklahoma	92	94
Tennessee	82	81
Texas	91	87
South Dakota	98	86
Alabama	96	80
Arizona	84	96
New Mexico	67	98
Pennsylvania	96	88
N =15	$\bar{x}_1 = \frac{1327}{15} = 88.47$	$\bar{x}_2 = \frac{1318}{15} = 87.87$

x₁ = Percentage of drivers involved in fatal collisions who were not distracted

x₂ = Percentage of drivers involved in fatal collisions who had not been involved in any previous accidents

$$H_0 : \mu_1 = \mu_2$$

H₁ : at least one mean is different

Standard deviation:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

$$S_1 = 9.95$$

$$S_2 = 6.94$$

Mean between samples:

$$\bar{\bar{x}} = \frac{88.47+87.87}{k=2} = 88.17$$

Standard deviation between samples:

$$s_{\bar{x}} = 0.42$$

Variance between samples:

$$ns_{\bar{x}}^2 = 15(0.42)^2 = 2.646$$

Variance within samples:

$$s_p^2 = \frac{(9.95)^2 + (6.94)^2}{k=2} = 73.583$$

$$\begin{aligned} F &= \frac{\text{variance between samples}}{\text{variance within samples}} = \frac{ns_{\bar{x}}^2}{s_p^2} \\ &= \frac{2.646}{73.583} \\ &= 0.037 \end{aligned}$$

Numerator and denominator degree of freedom:

$$\text{Numerator} = k - 1 = 2 - 1 = 1$$

$$\text{Denominator} = k(n - 1) = 2(15 - 1) = 28$$

Critical value with $\alpha = 0.05$ from F-distribution table:

$$F\text{-critical value} = 4.20$$

Conclusion: Since $F_{\text{test statistic}} < F_{\text{critical value}}$ ($0.037 < 4.20$), we fail to reject the null hypothesis. There is sufficient evidence to claim the percentage of drivers involved in fatal collisions who were not distracted have the same mean as percentage of drivers involved in fatal collisions who had not been involved in any previous accidents.

Conclusion

All in all, this project has given us some experiences in real life situation for gathering results and skills for us to understand how results can be obtain in various method such as through hypothesis 1 or 2 sample test, Correlation test, Regression, and Goodness of fit test. These data analysis proves the dataset that we choose which is in this case, bad drivers, can be proven from the initial hypothesis to either reject or fail to reject based on our initial assumption by using the level of confidence which can be depending on how confident we are. Each method has its different way of achieving results, but it all depends on what you are trying to gain. All the method's results are obtained from the test-statistic whether it is in the range of the critical value or not, we either reject or fail to reject the results. What is interesting about this finding is that in the Goodness of fit test, it proves that the percentage of drivers involved in fatal collisions who were not distracted in all 15 states, are all the same. On the contrary, there are also some data analysis results that seem to have either weak or no relation at all. From Regression test, the losses incurred by insurance companies for collision per insured driver does not depend on the number of drivers involved in fatal collisions per billion miles.

Appendix

Link for video presentation: <https://youtu.be/Ce8ZR3vVMPQ>

Source: **Bad Drivers** | Kaggle

Sample of raw data:

State	Number of drivers involved in fatal collisions per billion miles	Percentage Of Drivers Involved In Fatal Collisions Who Were Speeding	Percentage Of Drivers Involved In Fatal Collisions Who Were Alcohol-Impaired	Percentage Of Drivers Involved In Fatal Collisions Who Were Not Distracted	Percentage Of Drivers Involved In Fatal Collisions Who Had Not Been Involved In Any Previous Accidents	Car Insurance Premiums (\$)	Losses incurred by insurance companies for collisions per insured driver (\$)
Alabama	18.8	39	30	96	80	784.55	145.08
Alaska	18.1	41	25	90	94	1053.48	133.93
Arizona	18.6	35	28	84	96	899.47	110.35
Arkansas	22.4	18	26	94	95	827.34	142.39
California	12	35	28	91	89	878.41	165.63
Colorado	13.6	37	28	79	95	835.5	139.91
Connecticut	10.8	46	36	87	82	1068.73	167.02
Delaware	16.2	38	30	87	99	1137.87	151.48
District of Columbia	5.9	34	27	100	100	1273.89	136.05
Florida	17.9	21	29	92	94	1160.13	144.18
Georgia	15.6	19	25	95	93	913.15	142.8
Hawaii	17.5	54	41	82	87	861.18	120.92
Idaho	15.3	36	29	85	98	641.96	82.75
Illinois	12.8	36	34	94	96	803.11	139.15
Indiana	14.5	25	29	95	95	710.46	108.92
Iowa	15.7	17	25	97	87	649.06	114.47
Kansas	17.8	27	24	77	85	780.45	133.8
Kentucky	21.4	19	23	78	76	872.51	137.13
Louisiana	20.5	35	33	73	98	1281.55	194.78
Maine	15.1	38	30	87	84	661.88	96.57
Maryland	12.5	34	32	71	99	1048.78	192.7
Massachusetts	8.2	23	35	87	80	1011.14	135.63
Michigan	14.1	24	28	95	77	1110.61	152.26
Minnesota	9.6	23	29	88	88	777.18	133.35
Mississippi	17.6	15	31	10	100	896.07	155.77
Missouri	16.1	43	34	92	84	790.32	144.45
Montana	21.4	39	44	84	85	816.21	85.15
Nebraska	14.9	13	35	93	90	732.28	114.82
Nevada	14.7	37	32	95	99	1029.87	138.71

New Hampshire	11.6	35	30	87	83	746.54	120.21
New Jersey	11.2	16	28	86	78	1301.52	159.85
New Mexico	18.4	19	27	67	98	869.85	120.75
New York	12.3	32	29	88	80	1234.31	150.01
North Carolina	16.8	39	31	94	81	708.24	127.82
North Dakota	23.9	23	42	99	86	688.75	109.72
Ohio	14.1	28	34	99	82	697.73	133.52
Oklahoma	19.9	32	29	92	94	881.51	178.86
Oregon	12.8	33	26	67	90	804.71	104.61
Pennsylvania	18.2	50	31	96	88	905.99	153.86
Rhode Island	11.1	34	38	92	79	1148.99	148.58
South Carolina	23.9	38	41	96	81	858.97	116.29
South Dakota	19.4	31	33	98	86	669.31	96.87
Tennessee	19.5	21	29	82	81	767.91	155.57
Texas	19.4	40	38	91	87	1004.75	156.83
Utah	11.3	43	16	88	96	809.38	109.48
Vermont	13.6	30	30	96	95	716.2	109.61
Virginia	12.7	19	27	87	88	768.95	153.72
Washington	10.6	42	33	82	86	890.03	111.62
West Virginia	23.8	34	28	97	87	992.61	152.56
Wisconsin	13.8	36	33	39	84	670.31	106.62
Wyoming	17.4	42	32	81	90	791.14	122.04

Sample of process data:

State	Percentage Of Drivers Involved In Fatal Collisions Who Were Not Distracted (x_1)	Percentage Of Drivers Involved In Fatal Collisions Who Had Not Been Involved In Any Previous Accidents(x_2)
South Carolina	96	81
North Dakota	99	86
West Virginia	97	87
Arkansas	94	95
Montana	84	85
Kentucky	78	76
Louisiana	73	98
Oklahoma	92	94
Tennessee	82	81
Texas	91	87
South Dakota	98	86
Alabama	96	80
Arizona	84	96
New Mexico	67	98
Pennsylvania	96	88

