PROBABILITY AND STATISTICAL DATA ANALYSIS

SECI 2143

TITLE: PROJECT 2

SEMESTER : 2

SESSION: 2021/2022

LECTURER: DR. AZURAH A SAMAH

GROUP 5

NAME	MATRIC NO
AIN BATRISYIA BINTI NORAZLAN	A21EC0009
SITI NURKAMILAH BINTI SAIFUL BAHARI	A21EC0131
NASRUL AMIN BIN AB HADI	A21EC0099
QAISARA BINTI BADRUL HISHAM	A21EC0125

1. Introduction

In this project we are required to obtain a dataset and analyze and interpret it following the guide that is given. Based on this we have obtained a set of data from kaggle.com titled Amsterdam House Price Prediction of August 2021. We then identify its variable and interpret its result in a form of Hypothesis Testing, Correlation, Regression and Goodness to Fit Test. R Script was used in every test and calculation to get the desired output.

2. Scope and Objectives

- To find the equality of mean between House Price and House Area using two-sample test.
- To find that the House Area will effect the House Price or not.
- To predict the value of House Price by the value of the House Area.
- To find the relationship between House Area and House Price.

3. Datasets Used

Dataset URL:
https://www.kaggle.com/datasets/thomasnibb/amsterdam-house-price-prediction

Sample Size : 10

Variables:

I. House area (m^2)
- Measurement Level : Ratio
- Variable Type : Quantitative

II. House price (1000€)
- Measurement Level : Ratio
- Variable Type : Quantitative

House Area (m^2)	House Price (1000€)
32	225
78	325
105	475
128	580
138	720
145	1295
199	1625
258	1950
319	3925
497	4550

4. Descriptive Statistics

The dataset we have chosen for PSDA Project 2 is the Amsterdam House Price Prediction of August 2021. Our data is about the house price prediction in Amsterdam based on the variable house area which has a sample size of 10.

Variable	Data Type
House area (m²)	ratio
House price (1000€)	ratio

5. Inferential Statistic

Test	Description
Hypothesis testing	To test the equality of mean (two-samples-test)
Correlation	To find a positive correlation between house price and house area.
Regression	To predict the value of a House Price (1000€) based on the value of Area(m²).
Goodnss to fit test	To test the hypothesis that an observed frequency distribution fits some claimed distribution.

6. Interpretation of result

I. Hypothesis Testing

The purpose of hypothesis testing is **to test whether the null hypothesis can be rejected or approved**. In this project, hypothesis testing is carried out to test the equality of mean between house area and house price. We will be using two-sample test for this data.

By using R Studio,

```
        Welch Two Sample t-test

data:  dataset$`House Area (m²)` and dataset$`House Price (1000€)`
t = -2.85, df = 9.15, p-value = 0.019
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2469.23  -284.97
sample estimates:
mean of x mean of y
    189.9    1567.0
```

- **Null and alternative hypotheses.**

$H_0 : \mu_{price} - \mu_{area} = 0$

$H_1 : \mu_{price} - \mu_{area} \neq 0$

- **Level of significance.**

$\alpha = 0.05$

- **Criterion (critical values) for rejecting the null hypothesis.**

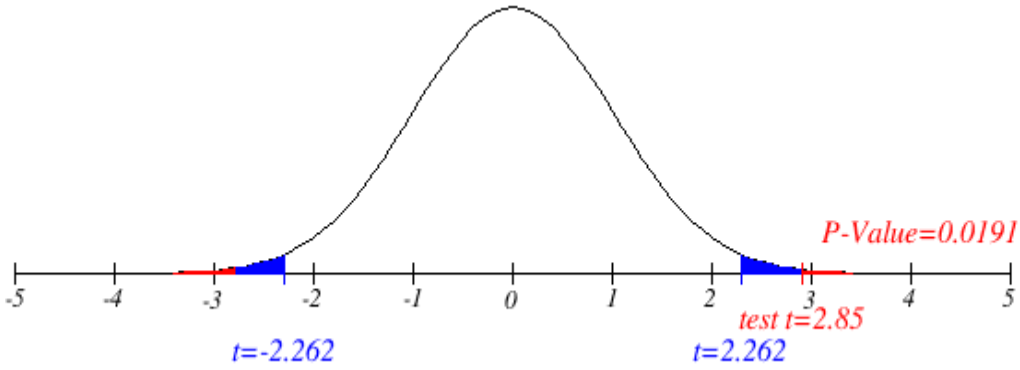
  From t-distribution table, the critical value is equal to ± 2.262

- **Compute the test statistic.**

$$t = \frac{\left(\overline{X}_{price} - \overline{X}_{area}\right) - \left(\mu_{price} - \mu_{area}\right)}{S_p\sqrt{\frac{1}{n_{price}} - \frac{1}{n_{area}}}}$$

$$t = 2.85$$

- **Decide whether to reject or fail to reject the null hypothesis.**



Reject null hypothesis.

- **Interpret results.**

Test statistic value falls inside the critical region. Therefore, reject the Ho.

There is enough evidence to support the claim that there are true difference in means between house area and house price which is not equal to 0.

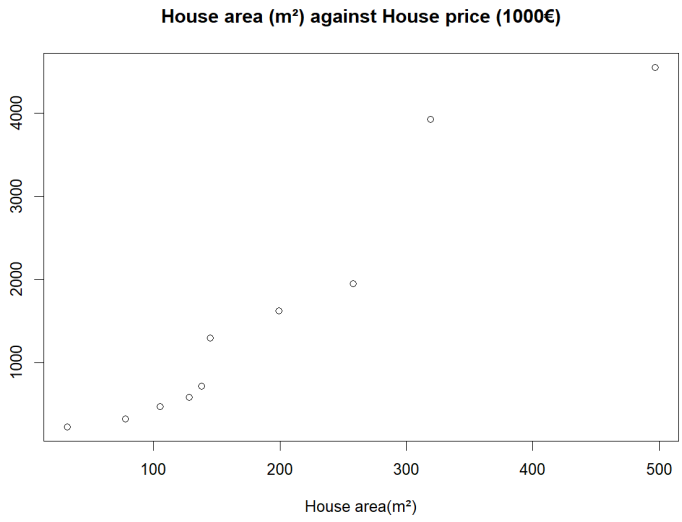
II. Correlation

This test is conducted to show whether there is a linear relationship between house area and house price

Variables :

x : House area (m²)

y : House price (1000 €)



From the scatter plot above, it indicates that there is a positive correlation between house area and the house price. It can be seen that the house price increases as the house area increases. However, there are a few outliers in the scatter plot.

By using R Studio we've got the value of,

1. Calculate sample correlation coefficient

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

r = 0.96205 , relatively strong positive linear association between x and y. The closer value r to 1, the stronger the positive linear relationship.

```
> cor(x,y, method = "pearson")
[1] 0.96205
```

2. Significance test for correlation

- Hypothesis Statement

  H_0 : $\rho = 0$ (No linear correlation)

  H_1 : $\rho \neq 0$ (Linear correlation exists)

- Test Statistic

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

  $t = 9.97$

- Significance level = 0.05, df = 8

- Critical region ,

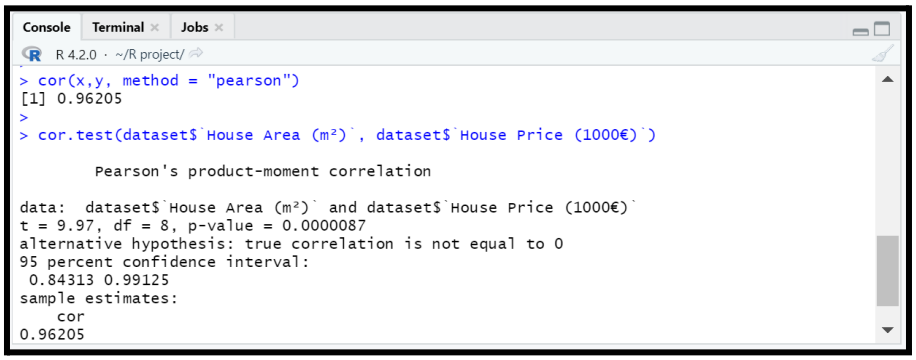
  $t_{0.025,8} = 2.306$

  $- t_{0.025,8} = 2.306$

❖ Decision : Since t = 9.97 > $t_{0.025,8}$ = 2.306, reject H_0

❖ Conclusion

Since t = 9.97 > $t_{0.025,8}$ = 2.306, reject null hypothesis. There is sufficient evidence of a linear relationship between house area and house price at the 5% level of significance.



III. Regression

Regression analysis is used to predict the value of the dependent variable based on the values of at least one independent variable. Since in our data the dependent variable is House Price (1000€) and the the independent variable is House Area (m²), then we can say that this test is to predict the value of House price (1000€) by the value of House Area (m²) in the Amsterdam. Regression is mostly used to determine the best equation to fit the data. A kind of regression analysis is linear analysis. y = a + bX is the linear equation, and y is the dependent variable.

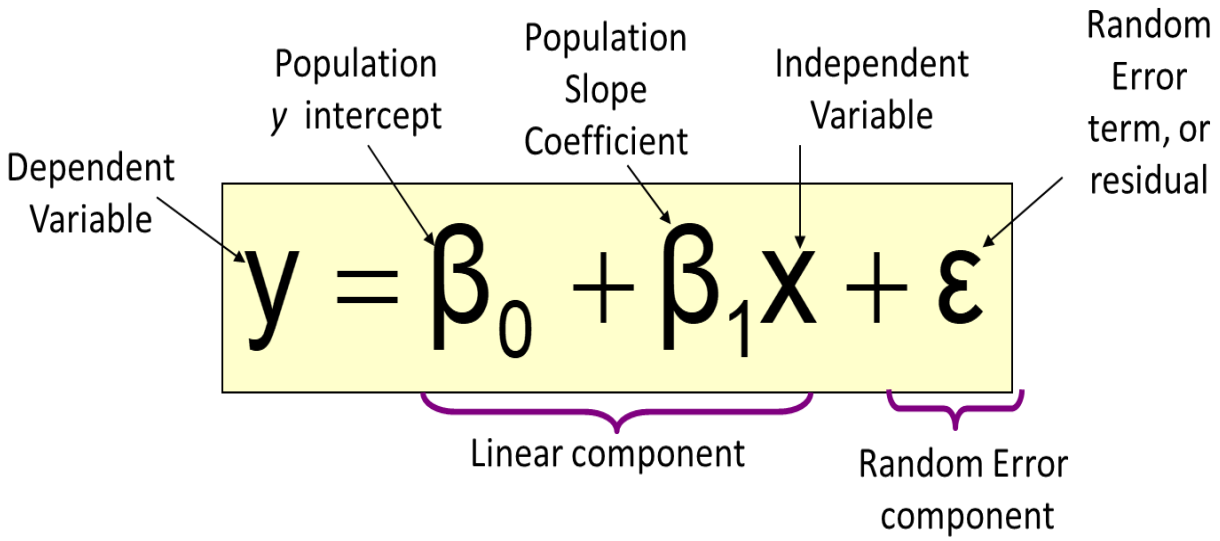
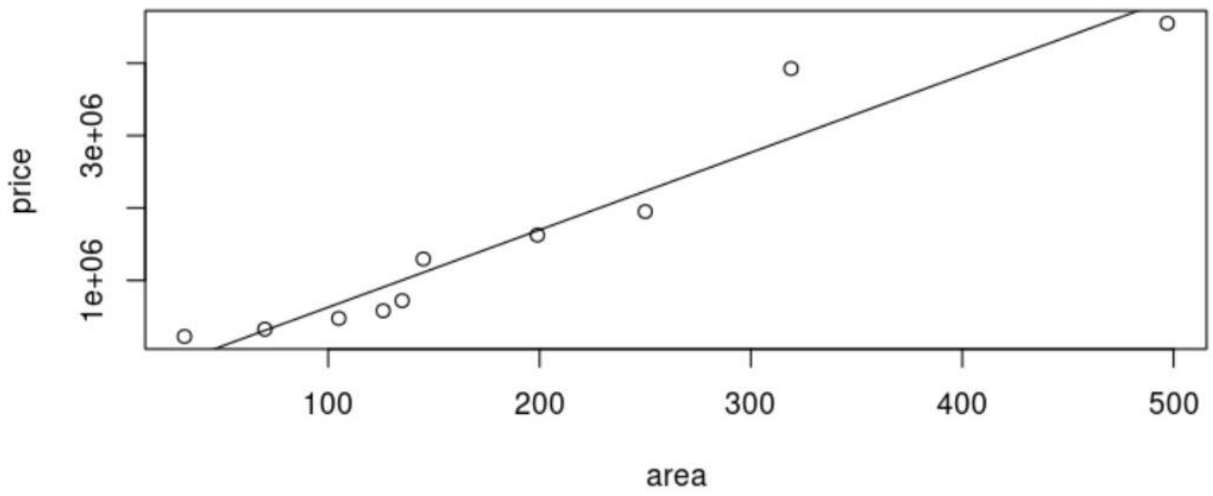


Figure of the scatter plot:



From the scatter plot above, it indicates that there is a positive linear relationship between the House Price and the House Area. It also shows the weaker relationship between the the House Price and House Area because we can see that some but not all of the variation in *House Price* is explained by variation in *House Area.*

```
Source
Console  Terminal  Jobs
R 4.2.0 · /cloud/project/
> area=c(32, 70, 105, 126, 135, 145, 199, 250, 319,497)
> price=c(225000,325000,475000,580000,720000,1295000,1625000,1950000,3925000,4550000)
> data=data.frame(area,price)
> View(data)
> names(data)
[1] "area"  "price"
> plot(area,price)
> model1=lm(price~area)
> abline(model1)
>
```

IV. Goodness to fit test

The goodness to fit test is used to test the hypothesis that an observed frequency distribution fits some claimed distribution. In this case it is used to determine whether the observed frequency which is House Price (1000€) is uniformly distributed.

By using R;

```
> # goodness to fit test
>
> housePrice <- c(225,325, 475, 580, 720, 1295, 1625, 1950, 3925, 4550)
>
> chisq.test(housePrice)$expected
 [1] 1567 1567 1567 1567 1567 1567 1567 1567 1567 1567
>
> expected <- 1567
>
> my.chi.stat <- sum((housePrice-expected)^2/expected)
> my.chi.stat
[1] 13344.01
>
> 1-pchisq(my.chi.stat,1)
[1] 0
>
>
> chisq.test(housePrice)

        Chi-squared test for given probabilities

data:  housePrice
X-squared = 13344, df = 9, p-value < 2.2e-16
```

- The above visual shows by hand method of the goodness to fit test as well as the simplified method.

- Hypothesis statement;

  null hypothesis : all probabilities are the same(uniformly distributed)

  alternate hypothesis : not all probabilities are the same(not uniformly distributed)

```
> my.chi.stat <- sum((housePrice-expected)^2/expected)
> my.chi.stat
```

- Under the influence of;

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- Due to the p value being a small number we can conclude that the probabilities are not equal therefore it is not uniformly distributed hence the null hypothesis is rejected.

7. Work Coordination

Group Members	Task
AIN BATRISYIA BINTI NORAZLAN	● Interpretation of Result ➔ Hypothesis Testing ● Conclusion
SITI NURKAMILAH BINTI SAIFUL BAHARI	● Dataset Used ● Interpretation of Result ➔ Correlation ● Video Presentation Editor
NASRUL AMIN BIN AB HADI	● Scope and Objectives ● Interpretation of Result ➔ Regression
QAISARA BINTI BADRUL HISHAM	● Introduction ● Descriptive Statistic ● Inferential Statistic ● Interpretation of Result ➔ Goodness to Fit Test

8. Reflections

Group Members	Reflection
AIN BATRISYIA BINTI NORAZLAN	I would say that this project is very challenging because this course is not easy to learn. It requires me to think outside of the box and I have to play with numbers and formulas to solve questions. I also struggle to use R Studio but somehow, I manage to deal with it by learning and doing a lot of researching. Therefore, I want to sincerely express my appreciation and thanks to my lecturer Dr. Azurah for being helpful and always guiding me to complete this project. She always gives ideas and would ask frequently about our progress throughout Project 2. I gained so much knowledge and experience from this project. There are a lot of new things that I knew when I was searching for information for my part.

Group Members	Reflection
NASRUL AMIN BIN AB HADI	For this project 2 of the course of Probability and Statistical Data Analysis, I can say that this project is actually pretty hard to complete. The most challenging part for me is the part of using the R studio and thankfully I did manage to complete the code even though I am facing some struggle in understanding the code or formula that I need to do. Plus, we manage to finish this project before the due date since we seem to help each other to ease the flow of our project. Next, I really want to thank to our course lecturer who is Dr. Azurah for being a very helpful lecturer and at the same time guiding us to fulfill the requirements of this project. Last but not least, I can conclude that this project did manage to make me gained more knowledge and information according to this course and I wish that I can use the knowledge in future.

Group Members	Reflection
SITI NURKAMILAH BINTI SAIFUL BAHARI	By finishing this project, I have gained a lot of new knowledge that I did not have previously. I gained more knowledge about R programming and statistics itself during my studies on our subject. I've finally figured out how to utilize R studio and apply statistical theory. For me, it is something completely fresh. Of course, the advice given by our professor, Dr. Azurah, on how to set up the R studio and the fundamental programming has been quite helpful to me. Not to mention, the encouragement from my group members keeps me motivated to finish my job. Since this is a group project, I need to give this project my all in order for it to be successful. After completing this project, I hope to learn more about R programming, statistical concepts, and how they can be applied in my future career. My understanding of this language needs to be updated and improved. To remain relevant in the field, I also need to upgrade both my soft and hard talents.

Group Members	Reflection
QAISARA BINTI BADRUL HISHAM	It would be an understatement to say that this project is hard. I feel that my biggest struggle is the use of Rstudio to solve the problem but after countless hours of researching ang input from team members and peers, I believe that I have got a clear picture of what needs to be done and hopefully have executed it according to the desired output. Thanks to our lecturer, Dr Azurah for her understanding and her leniency when it comes to giving us sufficient time to complete this project. I also appreciate my team members' helpfulness and teamwork throughout the process of completing this project.

9. Conclusion and Discussion

Summarize findings

Based on the analysis and interpretation that has been made, we can conclude that we reject the null hypothesis, that means there is enough evidence to support all the claim that there are true difference in means between house area and house price which is not equal to 0, there is a linear relationship between house area and house price at the 5% level of significance, there is a positive correlation between house area and house price and lastly, the probabilities are not equal therefore it is not uniformly distributed due to small p-value.

Discuss whether all objectives outlined in Section 2 have been achieved

From the report, we can said that all of the objectives that we outlined in Section 2 seems have been achieved. Since we can say that we manage to find the equality of mean between House Price and House Area which is not equal. There is a positive linear relationship between House Price and House Area. We also know that the House Area will effect the House Price. Next, we also get to predict the value of House Price by the value of the House Area using the regression method. Last but not least we also know that there is a relationship between the House Price and the House Area.

List of challenges faced throughout the study

- Difficulty in concentrating. Since we are facing online learning, it is likely to lose focus and experience a dramatic drop in productivity. We also find it is hard to focus because we had to sit in front of the screen for a long time which makes our minds wander.
- Experiencing low motivation. Since there are many things to do, with a lot of assignments and projects from other courses, we felt stressed and tired because we spend a lot of our time chasing deadlines and to balance commitments makes us struggling with time management.

Present improvement to be made in future

- Connect with classmate more often. Since we may start physical learning soon, we can meet and learn from one another, share challenges and concerns to overcome them together. Having peer support could help alleviate stress.
- Learn to adapt to project based learning and working in every aspects and situations. This is when organizational, collaborative, and time management skills can be use as basics in our further academic careers.

10. References

☐ Source of Data

Link: https://www.kaggle.com/datasets/thomasnibb/amsterdam-house-price-prediction

☐ References

Corporate Finance Institute. (2022, April 26). Regression Analysis. https://corporatefinanceinstitute.com/resources/knowledge/finance/regression-analysis/

DataCamp. (2018, July 18). *R Linear Regression Tutorial: lm Function in R with Code Examples*. Linear Regression in R Tutorial. Retrieved June 24, 2022, from https://www.datacamp.com/tutorial/linear-regression-R

☐ E- Portfolios

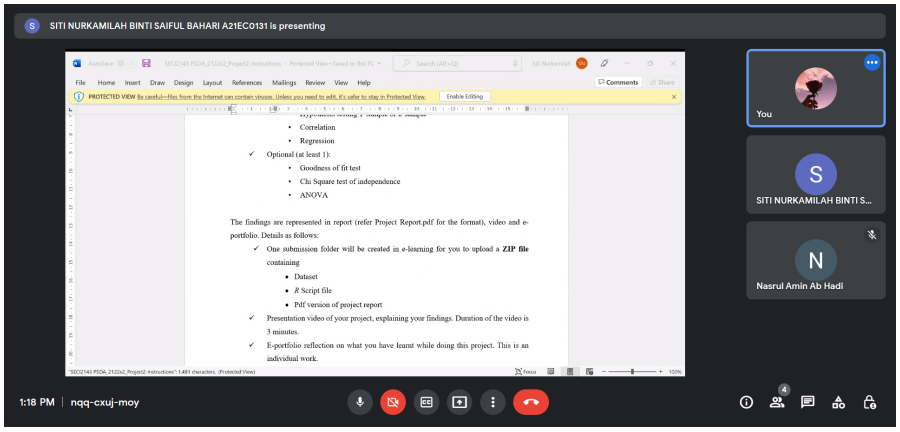
- ★ https://eportfolio.utm.my/user/ain-batrisyia-norazlan
- ★ https://eportfolio.utm.my/user/siti-nurkamilah-binti-saiful-b
- ★ https://eportfolio.utm.my/user/nasrul-amin-bin-ab-hadi
- ★ https://eportfolio.utm.my/user/qaisara-badrul-hisham

☐ Presentation video

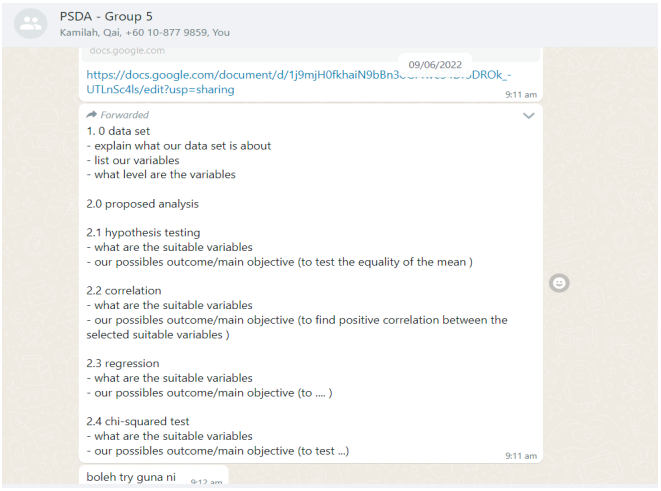
Link : https://youtu.be/QG9RV8GZwuE

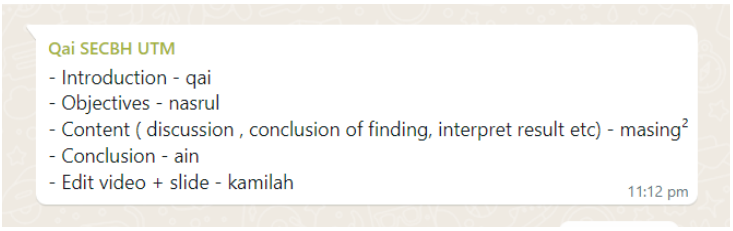
11. Appendices

Meeting/Discussion



Task Division





<< for *report*