CA2- INFERENTIAL STATISTICS

Corporate Flights

Marilia Meneses
Student ID: 2022048

Table of Figures

Table of tables

Table of graphics

Table of contents

1. Introduction

This report aims to answer all the questions of CA2- Integrated and it is divided into three sections: Research and Hypotheses Test, Correlation Analysis and Linear Regression Model.
The Research section focused on the Hypothesis test of a sample. A research need to be done in regards to find the mean of a population.
Secondly, I will carry out a correlation analysis between 2 variables from the data set. It will be check if correlation implies causation.
And finally I will be modelling a Linear Regression model and predicting information about these variables.
Data dictionary:

Column	Definition of Variable	Type of Variable	
		Qualitative/ Quantitative	**Categorical /Discr/Cont**
travelCode	travel id (primary key)	Qualitative	Categorical
userCode	user id (foreign key)	Qualitative	Categorical
from	travel from place	Qualitative	Categorical
to	travel to place	Qualitative	Categorical
flightType	flight type, first class, economic, premium	Qualitative	Categorical
price	price of the flight	Quantitative	Discrete
time	flight time	Quantitative	Continuous
distance	flight distance	Quantitative	Continuous
agency	travel agency	Qualitative	Categorical
date	travel date	Date	Date

Table 1 – Variables dictionary

The dataset analysed in this report is the same that was analysed in CA1- Descriptive Stats.
It simulates real corporate travel systems focusing on flights and hotels. It was provided by Argo Solutions available on the Kaggle website and contains 271888 observations and 10 variables whereas amongst them there are 6 qualitative variables and 3 quantitative variables and one date.
Context
Argo Solutions - A leading technology company in Latin America, developing solutions to facilitate expense management and corporate travel using technology as an enabler of these processes.

Given the fact that my background is in accounting, it would be interesting to analyse the expenses of corporate flights.

2. First Section: Research and Hypothesis Test

2.1 Research

ANAC (The National Civil Aviation Agency) is the Brazilian civil aviation authority, it has published reliable data about aviation and flights in Brazil. Due its relevance, I will conduct the hypothesis test based on a report from ANAC that says the average domestic plane ticket price in Brazil in October of 2019 was R$ 543.41 (ANAC, n.d.)

Figure 1 – Report from ANAC, October of 2019 (ANAC, n.d.)

2.2 Hypothesis test

A **hypothesis** is an assumption and a **hypothesis test** is a part of a statistical model that we are going to test this assumption and work with until we have sufficient evidence to reject it. (Spiegelhalter, 2019, pp.93, 94)

The objective of hypothesis test is by looking at a sample of data, we can test whether something is true or not and how confident we can be in the declaration we are making. (Lakin, 2011, pp.195, 199)

The two types of hypothesis testing are **null hypothesis** and **alternative hypothesis**. Null Hypothesis, usually denoted by H0, is the initial assumption. The alternative hypothesis, usually

denoted by H1, is that one that contradict the initial assumption and we want to prove. (Bruce, Bruce and Gedeck, 2020, p.93)

We want to find the minimum level at which the test rejects H0. It is only after calculating the **p-value,** which is usually within the range of 0 and 1, that we can accept or reject the null hypothesis. A smaller p-value means less confidence in our null hypothesis.

Hypothesis to be explored:

According to ANAC (The National Civil Aviation Agency) the average price of airline tickets in Brazil in October of 2019 was R$ 543.41. On calculating the average of the price on the same period, of 30 observations from the data set, the average turns out to be R$ 944.93.

The standard deviation is 358.80. I will analyse if the data provides sufficient evidence to reject the principal hypothesis or not at a 5% significance level.

The **significant level**, also represented by alpha α, is a risk of reject a null hypothesis when it is true. (Frost, 2019) In this case, means a risk of concluding that the average price for tickets is not R$ 543.41 when it is, actually.

We use **t-test** when we do not know the standard deviation of the population. (Lakin, 2011, pp.211) **Since I do not have the standard deviation of the population**, I will take a random sample of 29 observations from the data set to perform **a two tailed t-test in this case**, having the population mean represented by **μ,** the sample mean represented by $\bar{X}$ and the standard deviation of sample represented by **S**.

```
In [53]: #taking a sample of 29 observations from the data set, on the period of October,2019:

         october= (df['date'] > '2019-10-01') & (df['date'] <= '2019-10-31')
         filtered_df=df.loc[october]
         sample = filtered_df.sample(29,random_state=1)
         sample['price'].head()

Out[53]: 187960     857.32
         192230     544.86
         156148     744.11
         81727     1087.18
         64651      638.63
         Name: price, dtype: float64
```

Figure 2 – selecting a random sample of 29 observation from the data set.

In order to perform hypothesis testing, the null hypothesis will be called **H0** and the alternative hypothesis, **H1**.

So the hypothesis will be as follow:

- **Null hypothesis:** the average price of airline tickets in Brazil in October of 2019 was 543.41.
- **Alternative Hypothesis:** the average price of airline tickets in Brazil in October of 2019 was different of 543.41.

As the alternative hypothesis is **bidirectional** (H1 is different of H0) I will be running a **two-tailed** test using an arbitrated α level of 0.05.

Ho: $\mu = 543.41$

H1: $\mu \neq 543.41$

2.3 T – Test

I would apply Z-test in this case of study as I want to know if the mean of the sample is different from the population. However, since I do not have the standard deviation of the population, I will apply t-test to test the hypothesis, as both can be applied at the same use cases depending on the information we have available from the data set. (Howell, 2022)
Also according to Lakin (2011, pp.211): "When we don't know the standard deviation, the best we can do is to estimate it from the standard deviation of the sample we have taken, and then test with that."

T-test is one of the hypothesis test tool used to calculate the difference between the means of two groups. Using the t-test, we can test an assumption applicable to a population. (Hayes, 2021)
Stablishing the values to perform t-test on python:

```
In [57]: #stabilhing the values:

         sd = S/math.sqrt(29)
         alpha =0.05
         null_mean =543.41
         data = sample['price']

         # print mean and sd
         print('mean=%.2f stdv=%.2f' % (np.mean(data), np.std(data)))

         mean=936.21 stdv=331.38
```

Figure 3 – Stablishing the values to perform t-test

The mean of the sample is 936.21 and the standard deviation is 331.38.

Now, I will perform the test. In this function, I passed data, in value parameter. I passed mean value in the null hypothesis, in alternative hypothesis I checked whether the mean of the distribution of the sample is different than the population mean using parameter 'two-sided'. (docs.scipy.org, n.d.)

```
In [58]: #Now, I will perform the test:
         ttest_Score, p_value= stats.ttest_1samp(data, null_mean, alternative='two-sided') # stablishing 'two-sided'


         #Printting the results:
         print('\033[1m'+'t-test Score:'+'\033[0m'+'%.4f' % (np.array(ttest_Score)))
         print('\033[1m'+'p-value:'+'\033[0m'+'%.10f'% (np.array(p_value)))

         t-test Score:6.2722
         p-value:0.0000008825
```

Figure 4 – t-test and p-value scores

The t-test score is 6.2722 and the p-value is 0.0000008825. As the p- value is less than level of significant level, I will reject the null hypothesis.

```
In [59]: # We compare the p-value with alpha, if it is greater than alpha then we do not reject null hypothesis
         # else we reject it.

         if(p_value <  alpha):
           print('\033[1m'+"Reject Null Hypothesis"+'\033[0m')
         else:
           print(+'\033[0m'+"Fail to Reject NUll Hypothesis"+'\033[0m')

         Reject Null Hypothesis
```

Figure 5 – Reject Null Hypothesis

As conclusion, with 95% level of confidence, **I have enough evidence to reject my null hypothesis**. That means that the average price of tickets of flights in Brazil in October of 2019 was **different** from 543.41.

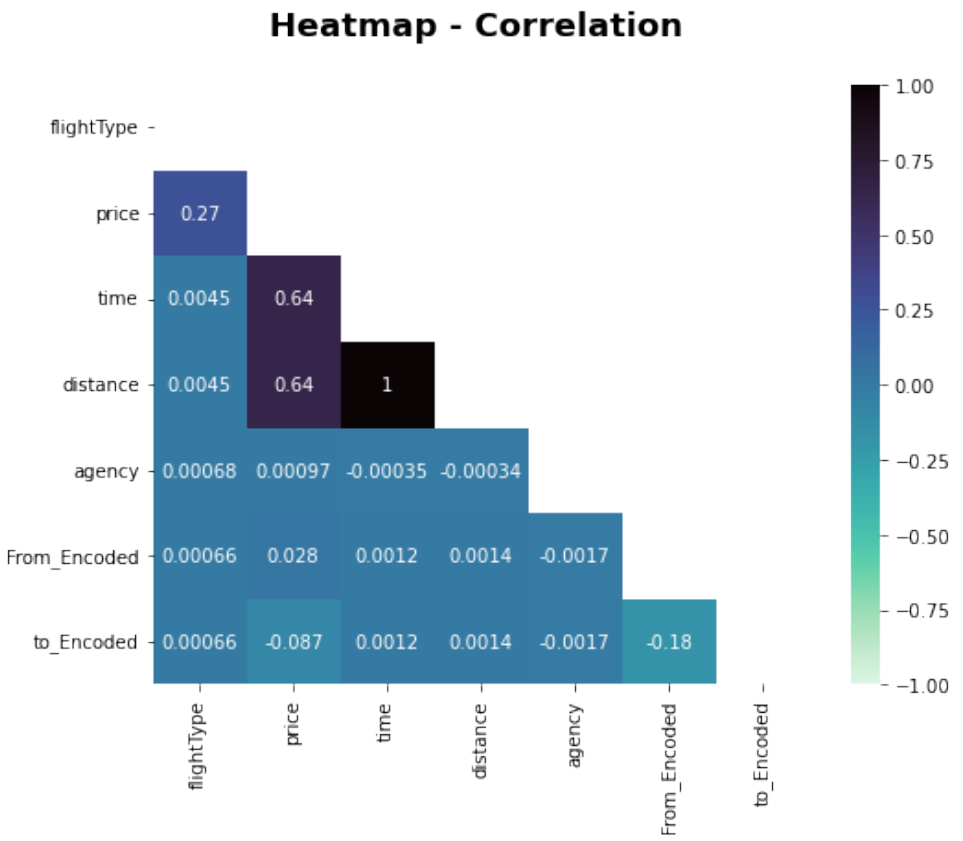
3. Second Section : Correlation

3.1 What is Correlation?

According to Haslwanter (2016, p.184) The coefficient correlation is a measure that tell us if two variables are related or not and if there is a relationship between them. Which means, if one variable changes, the other also changes.

In other words, correlation occurs when one variable can affect the value of another variable, or two variables may affect one another.
A correlation can be positive, when both variables move in the same direction or negative, when one variable's values moves up, the values of another variables moves down. (Brownlee, 2019)

Although the correlation measure the mutual relationship, most of the time correlation cannot imply causation. (Wijaya, 2021)



Graphic 1 – Heat map - Correlation

The heatmap above shows that there are a few variables with positive moderate correlation, for instance the variable ‘price, ‘distance’ and ‘time’. Others with very weak correlation and others with almost zero correlation. This information will be used to perform Linear Regression in two variables in the next topic.

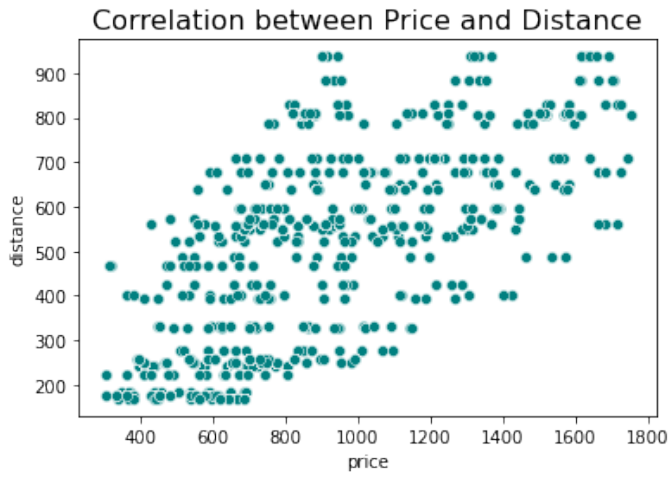
As informed previously, the heatmap showed how strong is the correlation among variables. The variables Price and Distance were chosen. Applying the correlation function, the result is 0.64, which means there is a moderate correlation between these two variables. It is clear, by looking at the scatterplot below, that there is a pattern between the variables.

```
In [62]: # Correlation matrix for variables chosen
         df[['price', 'distance']].corr(method='pearson')
```

Out[62]:

	price	distance
price	1.000000	0.641915
distance	0.641915	1.000000

Figure 6 – correlation matrix for variables chosen



Graphic 2 - Scatter plot – price and distance

Even though there is a moderate relationship between these two variables the Pearson Correlation can be small. (Wijaya, 2021)

In this case, even though a pattern can be seen between the variables ‘price’ and ‘distance’ when I plot a scatter plot, **the correlation between them does not imply causation because the**

variable price does not only depend of the variable distance. **I can not conclude that a small distance causes a low price of tickets because there are other variables to consider**, **which affect the price of the tickets for instance the variable 'FlightType'.**

3.2 Covariance

The covariance is a measure of the linear relationship tendency of the variables. This is a statistical method for analysing the behaviour of two random variables in conjunction. It explains how much X deviates from its mean when Y deviates from its own mean. Here it is how the covariance is calculated:

cov(X, Y) = (sum (x - mean(X)) * (y - mean(Y))) * 1/(n-1)

3.3 Pearson Correlation

One of the most used correlation coefficients during the data analysis process is the Pearson Correlation. The linear relationship between the two continuous variables is measured by the Pearson correlation and it has a value between 1 and -1. (Wijaya, 2021)

When the correlation coefficient is closer to value 1, it means there is a positive relationship between the two variables. In other words, if one variable increase another variable also increases. On the other hand, if the correlation coefficient is closer to -1 it means there is a negative relationship, which means if one variable increases another one decrease. (Wijaya, 2021)

4. Third Section: Linear Regression

Regression analysis is used to predict the value of a dependent variable based on the value of at least one independent variable. It explain the impact of changes in an independent variable on the dependent variable.

The Linear Regression model is used to predict continuous values based on the relationship between the quantitative variables. As was observed previously in the scatterplot that there is a linear relationship between the continuous variables and "the linear regression is fitting a straight line to a set of observations." (Kane, 2017, pp.133, 134)

To understand the pattern showed in the previously scatter plot, it was performed a n.polyfit with degree one to get the coefficients of the linear equation that reflects the correlation between Price and distance.

```
In [65]: flight_fit = np.polyfit(df.price, df.distance, 1)

         flight_fit

Out[65]: array([  0.37002606, 192.70182343])
```

Figure 7- linear regression coefficient

After performing that, the coefficients calculated were **a=0.37** and **b= 192.70**. Using these values, it is possible to create the Linear Regression Model:

y = 0.37x + 192.70.

price: x

distance: y

Graphic 3 -Relationship between time and distance

The method used above gave us the coefficients and the Linear Model. It is possible to perform prediction using it, but there is a library on python to do that. It will be performed below the Linear Regression using Scikit Learn.

The main objective of a model is to make predictions using new information applied to the model. To make prediction in a large dataset it is necessary to perform a test train split, so a part of the dataset can be used to compare values predicted with actual values. This way it is possible to get the accuracy of the model.

```
In [73]: # Defining variables X and y
         X = df['price']
         y = df['distance']

         # Performing Train Test Split
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=1234)

         # While performing I got an error telling me to reshape the variables, so I reshaped all of them
         X_train= X_train.values.reshape(-1, 1)
         y_train= y_train.values.reshape(-1, 1)
         X_test = X_test.values.reshape(-1, 1)
         y_test = y_test.values.reshape(-1, 1)

         # Checking the coefficients
         print('\033[1m'+'b: '+'\033[0m', lr_reg.intercept_)
         print('\033[1m'+'a: '+'\033[0m', lr_reg.coef_)

         b:  [193.29684217]
         a:  [[0.36945618]]

In [74]: lr_pred = lr_reg.predict(X_test)                    # Using Linear Regression to predict values for X_test
         R_square = r2_score(y_test, lr_pred)                 # Compare actual values to predicted values
         print('\033[1m'+'Coefficient of Determination: '+'\033[0m'+'{:.4f}'.format(R_square))

         Coefficient of Determination: 0.4156
```

Figure 8- Coefficient of determination

Performing Linear Regression using Scikit Learn showed very similar values for the coefficients. The Coefficient of Determination is 0.4146 (41.46%), which means that 41.46% of Total Sum of Squares is maintained by using this predictive model.

5. Conclusion

In. the first section, carrying out the hypothesis test I was able to determine if I would reject the null hypothesis or not. After calculate the t-test and the p-value we could see that with 95% level of confidence I could reject the Null Hypothesis. In other words, the average price of tickets of flights in Brazil in October of 2019 was **different** from 543.41.

In the second section, I could observe that, even though a pattern can be seen between the variables 'price' and 'distance' when I plotted a scatter plot, the correlation between them does not imply causation because the variable price does not only depend of the variable distance. I cannot conclude that a small distance causes a low price of tickets because there are other variables to consider, which affect the price of the tickets, for instance the variable 'FlightType'.

Finally, in the last section, performing Linear Regression using Scikit Learn showed very similar values for the coefficients. The Coefficient of Determination is 0.4146 (41.46%), which means that 41.46% of Total Sum of Squares is maintained by using this predictive model.

Suggestion for future research:
To perform a machine learning model and calculate correlation coefficients between all pairs of variables.

6. References

ANAC (n.d.). *Microsoft Power BI - ANAC*. [online] app.powerbi.com. Available at: https://app.powerbi.com/view?r=eyJrIjoiMjJjZjA3YTQtNjYwMi00NjZhLTg5NTUtMzRhODZlN2U0ZTc5IiwidCI6ImI1NzQ4ZjZlLWI0YTQtNGIyYi1hYjJhLWVmOTUyMjM2ODM2NiIsImMiOjR9&pageName=ReportSection7a8d3f66e2d8c1e70619 [Accessed 27 May 2022].

Brownlee, J. (2019). *How to Calculate Correlation Between Variables in Python*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/ [Accessed 21 May 2022].

Bruce, P.C., Bruce, A. and Gedeck, P. (2020). *Practical statistics for data scientists : 50+ essential concepts using R and Python*. Sebastopol, Ca: O'reilly Media, Inc, p.93.

docs.scipy.org. (n.d.). *scipy.stats.ttest_1samp — SciPy v1.7.1 Manual*. [online] Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_1samp.html [Accessed 23 May 2022].

Frost, J. (2019). *Significance level - Statistics By Jim*. [online] Statistics By Jim. Available at: https://statisticsbyjim.com/glossary/significance-level/ [Accessed 23 May 2022].

Haslwanter, T. (2016). *An introduction to statistics with python : with applications in the life sciences*. Cham: Springer, Cop, p.184.

Hayes, A. (2021). *How t-tests work*. [online] Investopedia. Available at: https://www.investopedia.com/terms/t/t-test.asp [Accessed 24 May 2022].

Howell, E. (2022). *Statistical T-Test Simply Explained*. [online] Medium. Available at: https://towardsdatascience.com/statistical-t-test-simply-explained-b510045d69e [Accessed 24 May 2022].

Kane, F. (2017). *Hands-on data science and Python machine learning : perform data mining and machine learning efficiently using Python and Spark*. Birmingham: Packt, pp.133, 134.

Lakin, S. (2011). *How to use statistics*. Harlow: Prentice Hall, pp.195, 199.

Spiegelhalter, D. (2019). *STATISTICS : the art of learning from data.* pp.93, 94.

Wijaya, C.Y. (2021). *What it takes to be correlated*. [online] Medium. Available at: https://towardsdatascience.com/what-it-takes-to-be-correlated-ce41ad0d8d7f [Accessed 25 May 2022].