UNIVERSITY OF EDINBURGH
Business School

CMSE11432: Principles of Data Analytics 22/23

FORM FOR THE SUBMISSION OF ASSESSED GROUP COURSEWORK

Title: <Applications and Discussion of Exploratory Analysis in the Group Project>

Examination Number: <B223193>

Word Count: <789>

I understand that mark is provisional until ratified by the Faculty Examination Board.

Applications and Discussion of Exploratory Analysis in the Group Project

When we have no idea about the general characteristics of the dataset, particularly before data cleaning, it is necessary to conduct exploratory research to assist enhance our understanding of our target problem, saving a great deal of time and cost in the early phase of the research. Exploratory research can even make a substantial contribution to designing subsequent procedures of data analysis (Singh, 2007). In some ways, exploratory analysis is more like inductive reasoning that approached the origin from the very end of the problem.

Therefore, exploratory analysis plays a vitally important role in facilitating the understanding of the entire dataset from a holistic view and provides a guideline to carry out the next step. For example, there are a few commonly used circumstances as follows:

1. In comparison with calculating quantiles, plotting a box plot is much simpler and more explicit to interpret the distribution of the data.

For instance, as Figures 1a and 1b show, drawing the distribution of price may be highly beneficial to detect and exclude any outlier accordingly, which in this case is the extremely high price.

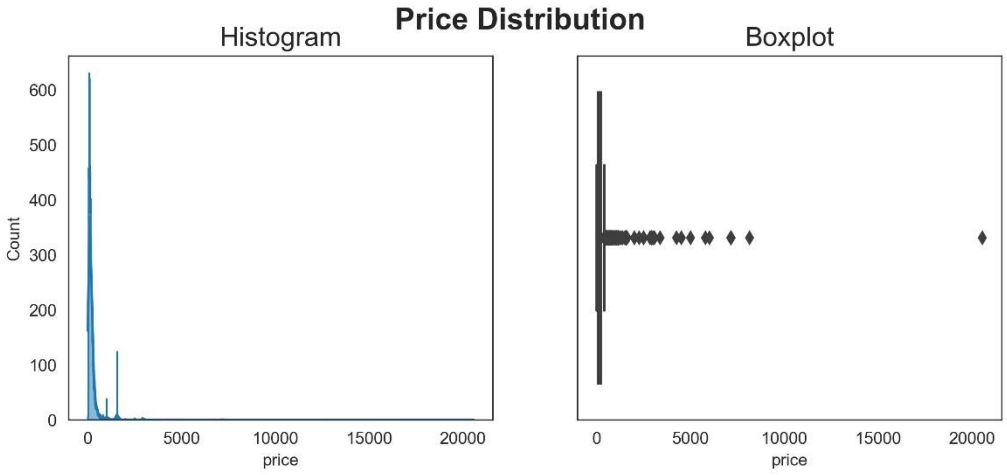

Figure 1a: Price Distribution (Extremely High Values Included)

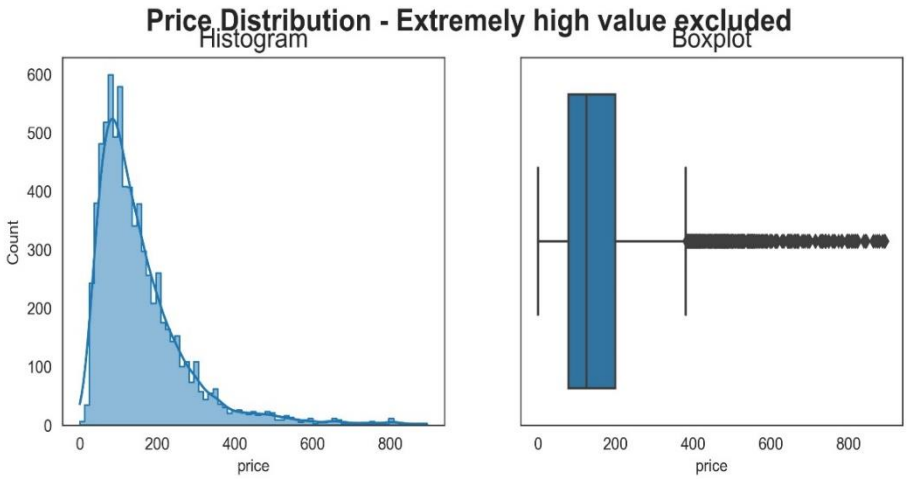

Figure 1b: Price Distribution (Extremely High Values Excluded)

2. A histogram can provide an intuitive view of the distribution of the population of the data to a great extent, which outweighs the complexity of interpreting a couple of statistics like *standard deviation* and *variability*.

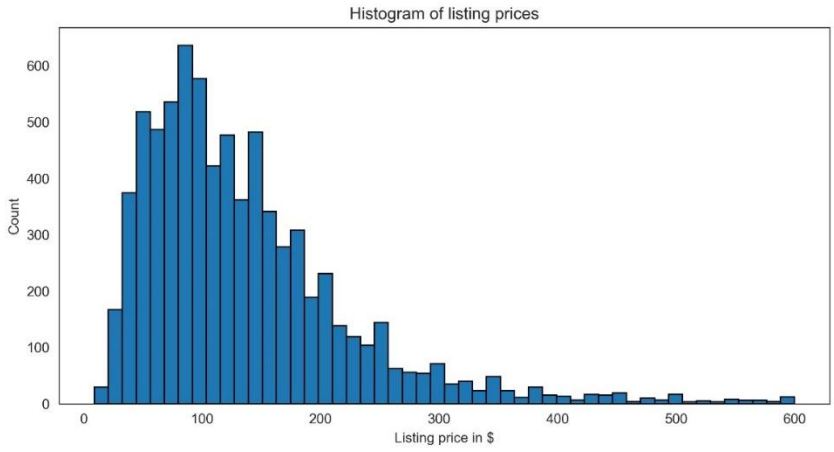

Figure 2: Histogram of Listing Price

For example, the histogram of listing prices is right-skewed in Figure 2, which indicates the mean should be larger than the mode of the listing price. As for skewed data, we need to do some processing to transform it to a normal distribution, and the common transformation methods involve logarithmic transformation and square root transformation (Norris and Aroian, 2004; Alexander, 2012).

3. When we draw a graph about the comparison between a wide variety of variables and detect a visible difference, it may direct us to the next procedure (Hypothesis testing) where analysts can further test and verify whether there is a significant difference regarding the variables of concern.

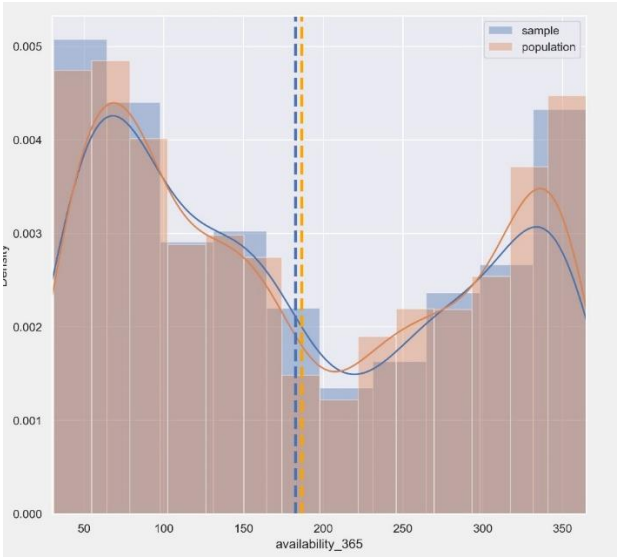

*Figure 3: Histogram of the Population and Sample Data for **Availability***

4. Before implementing the ANOVA to test the differences among variables' means, analysts must assure that the normality assumption has been met under type I error, in which case exploratory analysis will accelerate the test process as analysts can directly eliminate those that are visibly not satisfied.

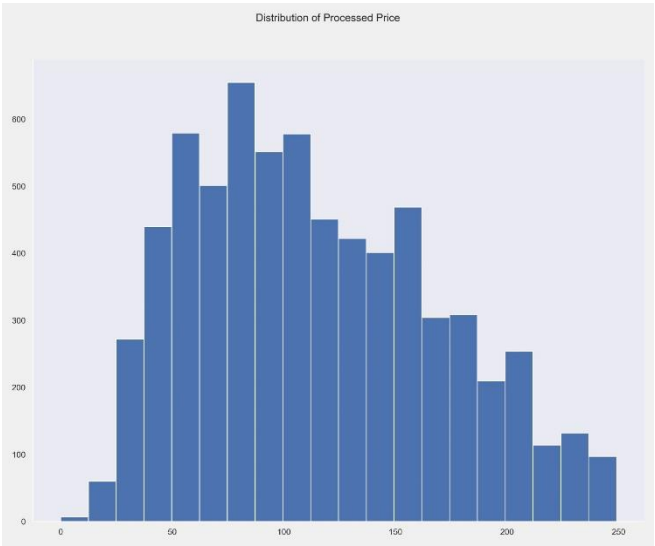

Figure 4a: Testing the Normality Assumption via Distribution Plot

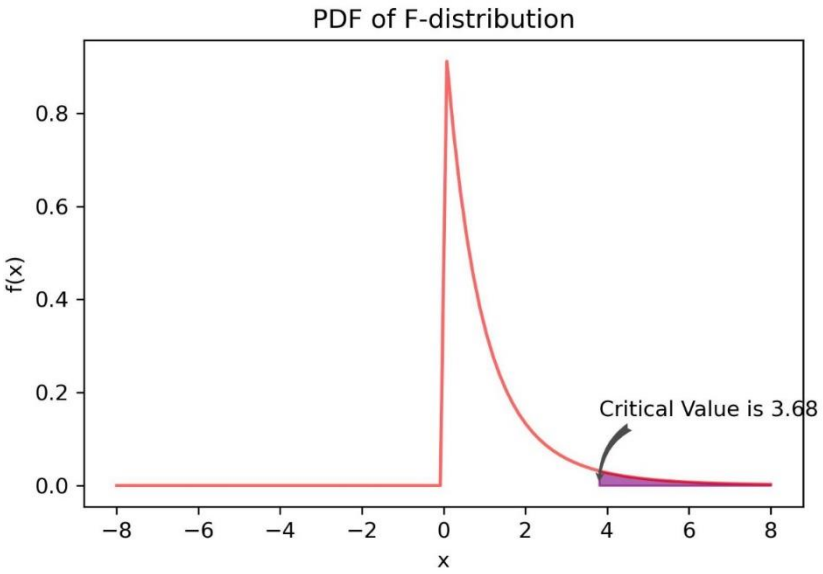

Figure 4b: Testing the Normality Assumption via Critical Region

5. As we detect an obvious positive or negative correlation between two or more variables by using a scatter plot or the "pair graph", it potentially gives us a hint that we may need to establish a linear regression model based on these variables with high relevance.

In addition, if the correlation coefficient is noticeably large, it can be concluded that there may exist a high statistical association between these variables of interest. More specifically, the other variable tends to change by a relatively fixed margin according to the change of one specific variable. It is slightly like the causational relationship whereas causation should imply a stronger bond, by which I mean that changes in one variable would certainly bring about the change in the other variable primarily because of the *"cause-and-effect" relationship*. And it should be emphasized that causation has far more stringent constraints than correlation and causation always imply a correlation between variables, not vice versa.

```
1  corr_matrix = df.corr()
2  corr_matrix['number_of_reviews'].apply(lambda x: abs(x)).sort_values(ascending = False).head(8)
```

```
number_of_reviews          1.000000
number_of_reviews_ltm      0.696445
reviews_per_month          0.364844
number_of_reviews_l30d     0.345905
id                         0.339574
host_id                    0.251254
maximum_nights_avg_ntm     0.176686
minimum_maximum_nights     0.169582
Name: number_of_reviews, dtype: float64
```

Figure 5: Eight Variables with the Highest Correlation Coefficients with ***Number of Reviews***

For instance, hereby to establish a simple linear regression first for the sake of simplicity, I inferred causality between the *number of reviews* and the *number of reviews over the last twelve months*, which is not robust and lacks sufficient validations to a certain extent. Therefore, other stronger evidence should be provided to demonstrate the rationality of the causality, such as Controlled Regression and Difference-in-Difference (Lechner, 2011; Aronow and Samii, 2016).

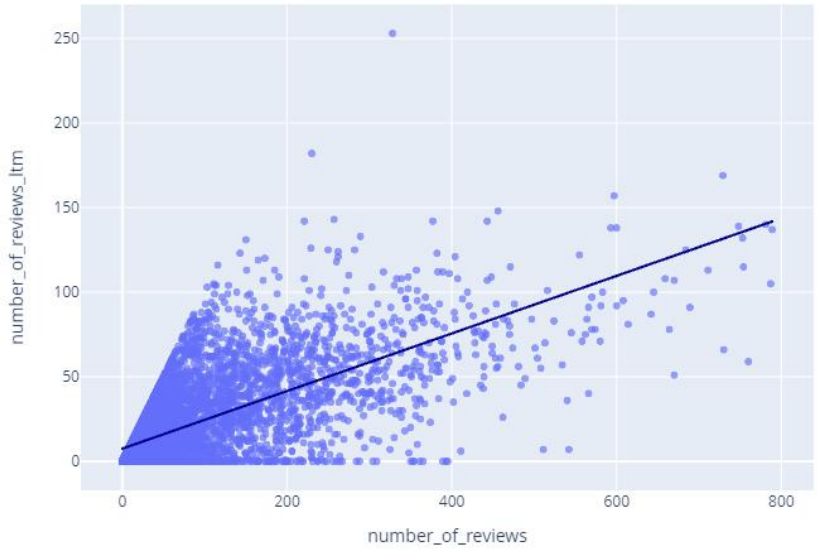

Figure 6: Linear Fit of two Highly-correlated Variables

More specifically, there are some typical concerns about whether we should construct a linear or multilinear regression model based on the high correlation between the target variable and those highly relevant independent variables, because it may raise confusion that it is these independent variables that lead to the change of our target. However, a couple of significant factors might be eliminated outside our model, for example:

- Analysts must create a viable hypothesis and have a clear objective when implementing exploratory analysis, otherwise, analysts would get overwhelmed by countless graphs full of useless information (Grove and Andreasen, 1982).

- The exploratory analysis could probably misdirect analysts when the dataset is highly biased (Whittingham *et al.*, 2006).

- Since it has to rely on the visualization of the data and most visualizations can only depict up

to three-dimensional graphs, the technique would be ineffective in dealing with high-dimensional data (Rauber, Merkl and Dittenbach, 2002).

Therefore, we must make it clear that all these factors mentioned above are necessarily taken into account and hence it will be more rational to make relevant decisions based on the model we have built.

In conclusion, exploratory analysis should be flexible and analysts ought to shift their direction according to new insights gained from the result of it (Saunders, Lewis and Thornhill, 2009).

REFERENCES

[1] Alexander, N. (2012) 'Review: analysis of parasite and other skewed counts', *Tropical Medicine & International Health*, 17(6), pp. 684–693. Available at: https://doi.org/10.1111/j.1365-3156.2012.02987.x.

[2] Aronow, P.M. and Samii, C. (2016) 'Does Regression Produce Representative Estimates of Causal Effects?', *American Journal of Political Science*, 60(1), pp. 250–267. Available at: https://doi.org/10.1111/ajps.12185.

[3] Grove, W.M. and Andreasen, N.C. (1982) 'Simultaneous tests of many hypotheses in exploratory research', *Journal of Nervous and Mental Disease*, 170, pp. 3–8. Available at: https://doi.org/10.1097/00005053-198201000-00002.

[4] Lechner, M. (2011) 'The Estimation of Causal Effects by Difference-in-Difference Methods', *Foundations and Trends® in Econometrics*, 4(3), pp. 165–224. Available at: https://doi.org/10.1561/0800000014.

[5] Norris, A.E. and Aroian, K.J. (2004) 'To Transform or Not Transform Skewed Data for Psychometric Analysis: That Is the Question!', *Nursing Research*, 53(1), pp. 67–71.

[6] Rauber, A., Merkl, D. and Dittenbach, M. (2002) 'The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data', *IEEE Transactions on Neural Networks*, 13(6), pp. 1331–1341. Available at: https://doi.org/10.1109/TNN.2002.804221.

[7] Saunders, M., Lewis, P. and Thornhill, A. (2009) *Research Methods for Business Students*. Pearson Education.

[8] Singh, K. (2007) *Quantitative social research methods*. Sage.

[9] Whittingham, M.J. *et al.* (2006) 'Why do we still use stepwise modelling in ecology and behaviour?', *Journal of Animal Ecology*, 75(5), pp. 1182–1189. Available at: https://doi.org/10.1111/j.1365-2656.2006.01141.x.