Project Phase 1

Statistical Inference
Spring 2020

University of Tehran
ECE Department

INTRODUCTION

In this project, we intend to study and analyze a series of real datasets with what you learned in this course. To begin analyzing a dataset, the first step is to get familiar with it. In the first step, this acquaintance can be made by observing the features of the dataset and distribution of the values and visualizing the data to make initial guesses about it. In the next step, by performing statistical tests, we make sure our guesses are correct and make our claims with certainty.

DATASETS

Based on the draw, you only have to work on one of the following datasets. For more information about your dataset, please refer to the mentioned references.

#	Name	Description
1	imdb.csv	The IMDB Movies Dataset contains information about 14,762 movies [1]
2	usedcar.csv	Dataset of used cars for sale with their technical specifications and prices [2]
3	insurance.csv	Dataset of health insurance costs of insurance contractors and their general information [3]
4	spotify.csv	Dataset of 19,000 Spotify songs and their details [4]

IMPORTANT NOTICES

- Use the R language in answering questions. Submit your codes in a separate file next to your report. **Reports without R codes are pointless**.
- In some datasets, you need to clear the data and convert the format and data type to more appropriate formats. So do this before answering the questions and explain the steps at the beginning of your report.
- If you need more categorical variables, you can add a new one to the dataset using some of your numerical variables. In this case, you need to describe the way you created the categorical variable from the numerical variable.
- In most of the questions, you should use the ggplot2 library to visualize and produce the desired charts.
- For each question, you need to fully explain your answer. An important part of the score will be attributed to your description. Drawing charts and performing calculations without sufficient explanations will result in losing the score. These descriptions show how much you understand the dataset. If you see interesting things in the diagrams, don't forget to mention them.
- When performing statistical tests, be sure to check the requirements for that test and write it down in your answer.

QUESTION 0

With answering these questions, you will get valuable information about your dataset:

a. Briefly describe your dataset and why studying your dataset can be interesting?
b. How many variables (features) and cases does your dataset have?
c. Is there any missing value in your data? Provide a summary on portion of missing values for each variable (feature) and describe how you handle these missing values for each variable (on what basis).
d. Using this elementary view of your dataset, which variables do you think may be the most relevant (contain some important information)? Why?

QUESTION 1

Choose one numerical variable from your dataset:

a. Plot the quantiles of this variable against the standard normal distribution. Discuss on the distribution of it and explain the result.
b. Categorize this variable into four intervals and Plot a pie chart that visualizes frequency of these four categories. Your chart should be colorized and the labels should contain each category with its percentage.
c. Plot a histogram for this numerical variable with an appropriate bin size.
d. Visualize density for this numerical variable.
e. Describe modality and skewness (calculate skewness).
f. Calculate mean, variance, standard deviation, and skewness.
g. Draw the boxplot, determine the upper and lower quartiles, whiskers, and the IQR.
h. What are outliers in this variable? Determine the outliers and their quantity.

QUESTION 2

Choose a categorical variable:

a. Plot the barplot for this variable.
b. Horizontal bar plot sorted by frequency
c. Create a frequency table for this variable.
d. Plot a violin plot for this variable.

QUESTION 3

Select two numerical variables and use the gplot (ggplot2) to answer the following questions:

a. Draw the scatter plot for these two variables and describe the relation between them and explain this relation if you can.
b. Select a categorical variable, and determine the samples either by the symbol or by the color (or by the both) in scatter plot that has been drawn in section "a".
c. Calculate the correlation coefficient for these two variables. Using the "cor.test" function, we can also test the significance of a correlation. Are the variables correlated? According to the test, what is shown by the p-value and what is intuition of p-value?
d. A **hexbin plot with marginal distribution** is like a two-dimensional histogram. The data is divided into bins, and the color strength of each bin represents the number of data points in that bin. Also, each dimension has its own distribution in front of its axis. Draw the hexbin plot with marginal distribution for chosen variables. What is your interpretation? Discuss the bin size and how it changes the result.

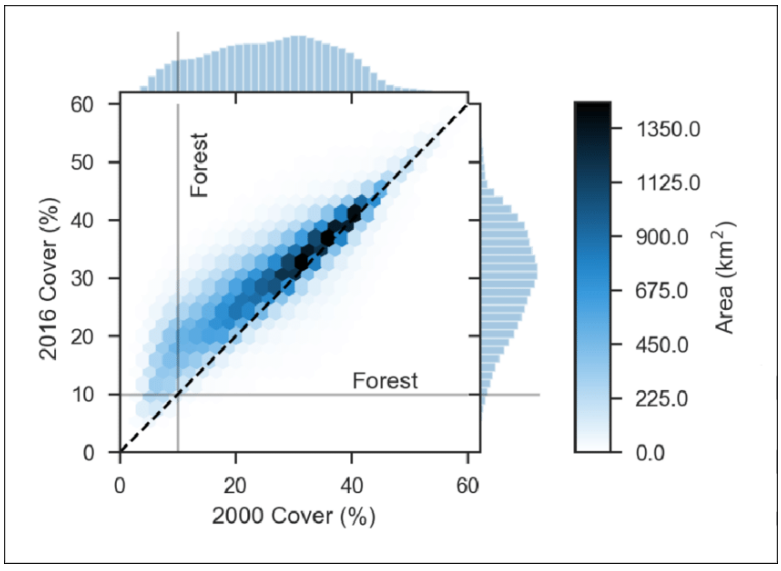

Figure 1. Hexbin (2D histogram) and marginal histograms of canopy cover in 2000 and 2016. Forest is defined as areas with at least 10% cover.

e. Draw the 2D density plot for chosen variables. How do you interpret the resulting graph? Describe the advantages and disadvantages of the 2D density and hexbin graph.

QUESTION 4

Consider a group (more than 3) of numerical variables from your dataset.

a. Display all the bivariate relations between the variables using a correlogram[1] where each element is a scatter-plot between two variables.
b. Describe the relations between the variables. Can you find any interesting pattern between them?
c. Create a heatmap correlogram from your variables. Annotate each cell with their corresponding Pearson's correlation coefficients. Use red for positive correlation and dark blue for negative correlation.
d. Choose 3 numerical variables from the group, Draw the 3D-scatterplot for these variables. Describe the relation between them.
e. Select a categorical variable from the dataset and show the categorical variable values as the color of the samples **with a bar** in the 3D-scatterplot that be drawn in section 'c'.

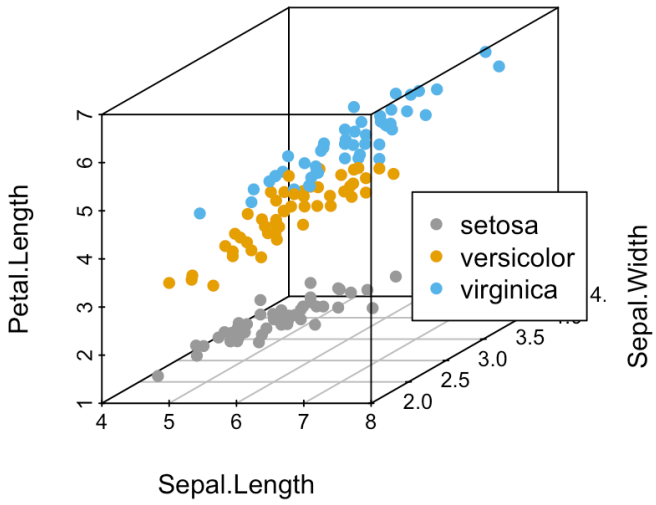

Figure2. 3D-scatterplot with a categorical variable coloring

[1] https://www.r-graph-gallery.com/correlogram/

QUESTION 5

For each below chart types; consider two categorical variables from your dataset that could be better described than others and then draw the chart:

a. Contingency table
b. Grouped bar chart
c. Segmented bar plot
d. Mosaic plot

QUESTION 6

Choose a numerical variable in your dataset:

a. Calculate a 98% confidence interval for the mean of this variable.
b. Interpret this confidence interval. In this context, what does a 98% confidence level mean?
c. Plot a Bar-plot for this variable with its confidence interval.
d. For the mean value of this numerical variable, design a hypothesis test and by finding the p-value, confirm or reject your assumption. Interpret this p-value.
e. Calculate a 95% confidence interval for this numerical variable. Based on this confidence interval, do these data support the hypothesis that you have designed? Explain it.
f. Calculate type II error.
g. Calculate the power and Explain the relationship between the power and effect size.

QUESTION 7

In this question, you will conduct a hypothesis test for two numerical variables. Choose a random sample of 25 data points from the dataset and choose two numerical variables that are not of a corresponding quantity. We would like to use this data to compare the average quantity between the two variables.

a. Should we use a t-test or z-test? Explain it.
b. Design a hypothesis test to see if these data provide convincing evidence of a difference between mean values. Does the result agree with the 95% confidence interval?

QUESTION 8

Choose a numerical variable that has outliers and we cannot apply CLT based methods we have learned so far.

a. Calculate a 95% confidence interval for the median of this variable using percentile method.
b. Calculate a 95% confidence interval for the median of this variable using standard error method.
c. Is there any difference between these two calculated confidence intervals? Explain your reasoning.

QUESTION 9

Answer this question based on the dataset assigned to you. This requires that you verify your answer by performing statistical tests and satisfying the required conditions. Perform random sampling if necessary. Drawing a chart to clarify your answer has an extra score.

- IMDB Movies:

  Do movies with more news articles get higher rankings in IMDB? Is there sufficient evidence to support this?

- Used Car:

  Are used cars with an automatic gearbox higher in the price? Is there enough evidence to prove this?

- Insurance:

  Is there enough evidence to claim that having more children increases medical costs?

- Spotify:

  Are longer songs on Spotify more popular? Is there enough evidence to prove this?

REFERENCES

[1] https://www.kaggle.com/orgesleka/imdbmovies

[2] https://www.kaggle.com/irfanazeem/used-cars-sale-price#car_train.csv

[3] https://www.kaggle.com/annetxu/health-insurance-cost-predicition

[4] https://www.kaggle.com/edalrami/19000-spotify-songs#song_data.csv