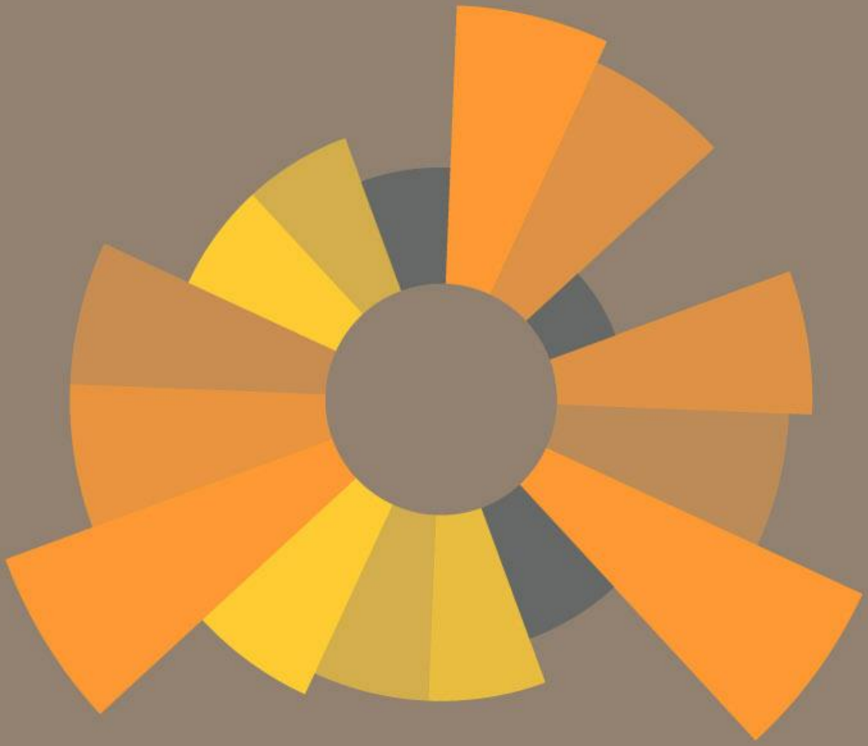
Statistical Inference | Spring 2018

Project Phase 1



University of Tehran

ECE Department

Esfand 1396

INTRODUCTION

One of the following datasets is assigned to you. Explain why studying the given dataset can be interesting. Describe the types of variables existing in the dataset. By doing this simple task, you gain an initial understanding of your dataset. Knowing your dataset is the first step in data science and it usually works as a helpful step for more advanced analyses.

Using this elementary view of your dataset, what do you think about the variables that might be relevant (contain some important information)? Why? (Note that in this section, we only want you to express your intuition about the relationship between the variables without doing any quantitative analysis.)

DATASETS

Dataset Name	# Obs.	# Vars	Description
DOTA2	42k	53	Dataset of DOTA2, popular multiplayer online battle-arena game • Original dataset contained records of 50k matches. Some data cleaning has been done (such as deleting matches with less than 10 players) which is resulted in 420k records of 10-player matches. A 10% sample of matches is taken which includes 42k instances overall. • Features description: **DOTA2_Desc.xlsx** • You can find more about the dataset at **[A]**
FIFA18	18k	185	The dataset of all soccer players contained in FIFA 18 • You can find more about this dataset at **[B]** • Dataset is scraped from **[C]**
Terrorism	56k	134	Dataset of details on terrorist attacks around the world since 1995 up to 2016 (with at least 1 killed person) • You can find more about the dataset here **[D]** • Codebook of variables is available at **[E]** - page 12
BikeSharing	18k	16	Dataset of hourly count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information • Find more about the dataset and its features at **[F]**

TOOLS

In most of the questions it is necessary to use **ggplot2** to generate the requested charts and plots. You can find more about this elegant visualization package in references [1], [2].

We also highly recommend using **data.tables** package for data manipulation. It is a high performance in-memory data processing library with an intuitive and easy to catch syntax (especially if you have experience using **pandas** library of python). You can learn how to use data.tables in [3], [4].

Important Requirements

For each question you answer, you have to fully explain the **meaning** of your analysis and interpret the generated plot and what you observe. The more reasonable your analysis is, the more positive effect it has on acquired grade of the corresponding question.

Furthermore, whenever you need to do hypothesis testing you **must check** all of the pre-requisite conditions (such as sample size, skewness etc.). Finally validity of results should be discussed.

Question 0.

Does your dataset have missing values? 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).

Question 1.

Select a numerical variable from your dataset. Use qplot (ggplot2) to answer the following questions:

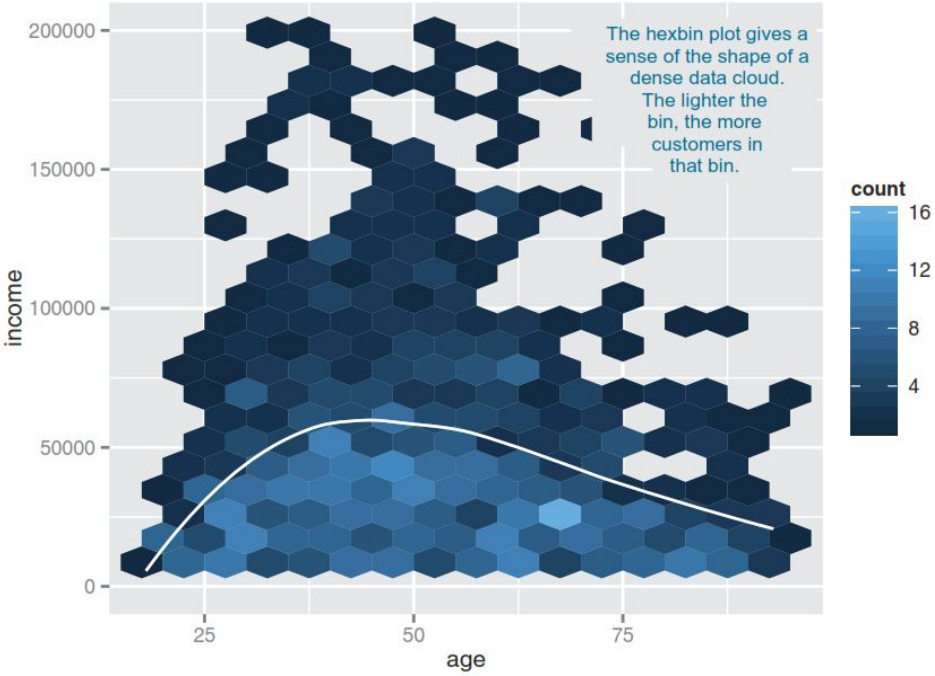
1. Plot the histogram of this numerical variable with an appropriate bin size.
2. Visualize density for this numerical variable.
3. Describe modality and skewness.
4. Calculate mean, variance, standard deviation, and skewness.
5. Plot boxplot, determine the upper and lower quartiles, IQR, lower inner fence and upper outer fence.
6. What are outliers in this variable? Determine the outliers and their quantity.

Question 2.

Consider two numerical variables and use qplot (ggplot2) to answer the following questions:

1. Draw the scatterplot of these variables. Describe the relationship between them.
2. Calculate the correlation coefficient between the chosen variables.
3. If you have a scatterplot with many data points, it can be hard to see the trend shown by the data. In this case you may want to add a smoothed line (fitting curve) to the plot. In this section, you should add a smoothed line and compare your result with the previous section.
4. Repeat parts 1-3 of this section for **two** other pairs of numerical variables (you may reuse a variable, but the pairs must be different). You will have 3 scatterplots drawn overall.

5. A **hexbin** plot is like a two-dimensional histogram. The data is divided into bins, and the number of data points in each bin is represented by color or shading. An example of such a plot is given below. Draw a hexbin plot with a fitting curve for your chosen numerical variables. How do you interpret the resulting graph? Discuss the bin size and how it changes the result.



The hexbin plot gives a sense of the shape of a dense data cloud. The lighter the bin, the more customers in that bin.

Figure 1. Hexbin-plot of income vs. age with a smoothing curve in a specific dataset

6. Take ***6* numerical variables** of your dataset (or all numerical variables if your dataset has less than 6 of them). Create a correlation heatmap (matrix) of them. Larger correlation values must have hotter color for their corresponding matrix element. Use <u>red</u> for <u>positive</u> and <u>dark blue</u> for <u>negative</u> correlation. A sample correlation heatmap is given below:

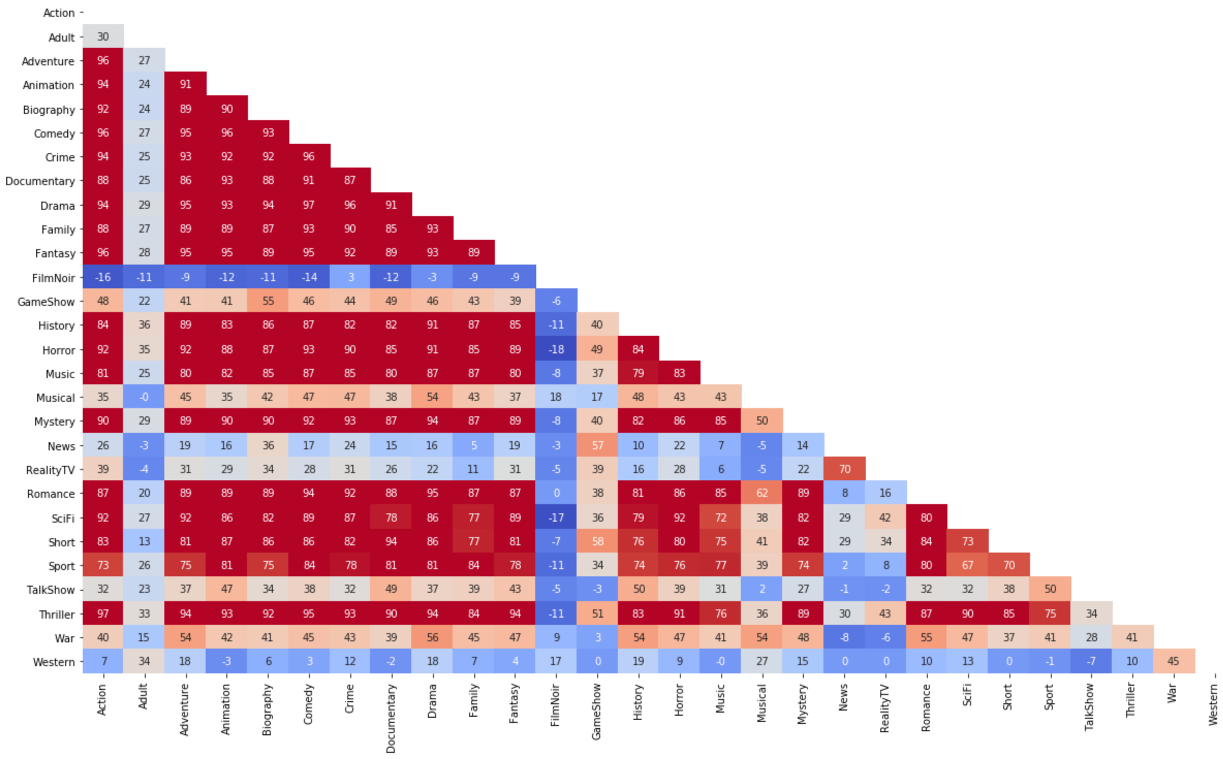


Figure 2. An example of a correlation heatmap. Correlation values has been multiplied by 100. Each variable is a movie genre which takes values 0, 1.

Question 3.

Choose a categorical variable. For this variable plot:

1. Bar plot
2. Horizontal bar plot sorted by frequency
3. Frequency table

Question 4.

Select two categorical variables and plot:

1. Contingency table.
2. Segmented bar plot
3. Side-by-side bar chart
4. Mosaic plot

Question 5.

When a dataset includes a categorical variable and one or more continuous variables, you will probably be interested in knowing how the values of the continuous variables vary with different levels of the categorical variable. Box plots and jittered points offer two different ways to do this. Consider the **diamonds** dataset of ggplot2 library:

1. Select a sample of size 200 from this dataset randomly. Use qplot (ggplot2) and explore how the distribution of price per carat varies with the color of the diamond using jittering and box-and-whisker plots.
2. What are the strengths and weaknesses of each of these two methods?
3. What is your interpretation of these plots?
4. Plot log transformation of variables: price vs. carat.

Question 6.

Choose a numerical variable in your dataset:

1. Calculate a 98% confidence interval for the mean of this variable.
2. Interpret this confidence interval.
3. For the mean value of this numerical variable, design a hypothesis test and by finding the p-value, confirm or reject your assumption.
4. Calculate type II error.
5. Calculate the power and justify that calculated value.

Question 7.

Choose a numerical and a categorical variable with more than two levels. Divide the observations of your dataset into different groups such that each group represents a level of the chosen categorical variable, i.e. all observations of a group must have the same level for this variable.

1. Using the ANOVA test, compare the mean value of the numerical variable in the groups.
2. Choose two of the groups, perform a hypothesis test for the mean difference of the selected numerical variable in these groups and calculate the p-value. Make a decision and explain the result using a significance level of 5%.

REFERENCES

1. Intro to Data Visualization with R & ggplot2: https://youtu.be/49fADBfcDD4
2. ggplot2: Elegant Graphics for Data Analysis (Use R!)
3. https://www.datacamp.com/courses/data-analysis-the-data-table-way
4. http://r-datatable.com

DATASET-RELATED LINKS

A. [DOTA2]: https://www.kaggle.com/devinanzelmo/dota-2-matches
B. [FIFA18]: https://www.kaggle.com/kevinmh/fifa-18-more-complete-player-dataset
C. [FIFA18]: https://sofifa.com
D. [Terrorism]: https://www.kaggle.com/START-UMD/gtd
E. [Terrorism]: http://start.umd.edu/gtd/downloads/Codebook.pdf
F. [BikeSharing]: http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset