Applied Statistical Methods
Statistics GU3105 — Fall 2021

Homework 2

The following problems are due on Tuesday, October 12, 11:59pm.

1. Continue with the .csv file you obtained in HW1.

   (a) Fit a regression and analyze the output. Fit a simple linear regression model to the scatter plot in HW1 and report the fitted slope and its p-value. Why is the slope relevant to our problem here (at most 3 sentences)?

   (b) Which assumptions were required for your p-value in 1(a) to make sense?

   (c) Test whether the inflation is independent of the unemployment rate via permutation test Please write the code for the following:

   - Calculate the correlation between the 2 variables and record this value.
   - Repeat the following 1000 times: shuffle the order of one of the variables, then recalculate the correlation. In other words, we're making the two variables independent from one another.
   - Please plot the histogram of these recalculated correlations against the original correlation.
   - Please comment on what you can infer from the histogram

2. In this question, you will look into the comments associated with various NYTimes articles.

   (a) There's the comments data in the file `nytimes_2020_articles_with_comments.json`. Load this using `jsonlite::read_json()`. What are the features of the data?

   (b) What is the main headline of the article with the most comments?

   (c) Please process the comments data into a data frame where each row is a different comment, and the columns each contain:

   - the number of recommendations received;
   - the displayed name of the commenter on NYTimes;
   - the update time;
   - the approved time;
   - whether the article was selected by the editors;
   - the number of words in the comment (splitting the comment by spaces is sufficient);
   - the number of unique words in the comment;

- the rank of the update time (e.g. the first comment would have rank 1, second would have rank 2, etc).

(d) If our population of interest was all the news articles from NYTimes in March 2020. What type of sample would you say we have?

Project Question Series 1

Background

Now that everywhere reopens and you are able to travel for vacations, it is time to see if your flights will be on time. **Notice that below are "thought exercise" before seeing the real data.**

1. What kind of data you can use? How are you going to collect them?
2. What potential issue you might face with the dataset you choose?

Dataset

You are going to use a part of dataset `pnwflights14` provided on Coursework. (Hint: Google `pnwflights14` and see what you find!)

1. Familiarize yourself with the data, including
   - How many columns are in the dataset? What does each column stand for?
   - How many airlines are in the dataset? Please aggregate the number of flights per month and visualize this with clear labels.
   - What data quality issues have you noticed? What results will be impacted by these issues?
   - Do certain airlines departure/arrive late their flights more often than others?
   - Are the arrival delays independent of the destination airport? How would you test this idea?
   - Is there a cyclical pattern to the arrival time? Please use a graph and a paragraph to justify your answer.
   - Are arrival delays the same across airlines (i.e. carriers)?
2. Do you realize something not quite right with the dataset? Please state which cleaning/pre-processing steps you are going to apply, and clean the dataset accordingly.