National University

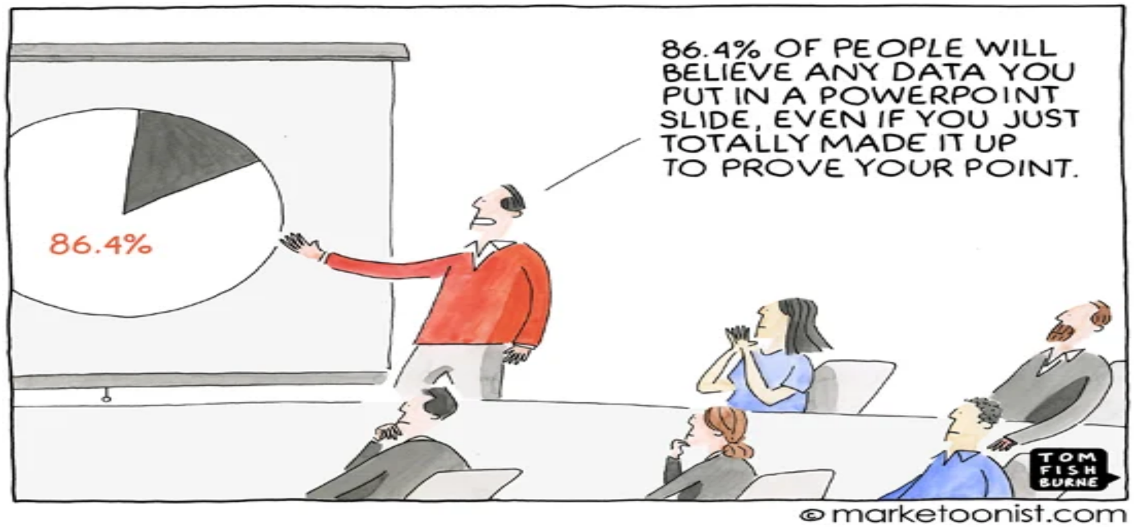
Of Computer & Emerging Sciences Chiniot- Faisalabad Campus

CS 4048 – Data Science (Fall 2025)

Assignment 3

Topics Covered: ▪ Basic EDA & Inferential statistics	**Submission Deadline:** *Saturday – November 15, 2025 by 23.59 sharp*
Submission Guidelines: ▪ Prepare a Python notebook (.ipynb) for each problem separately (code + textual responses) and submit a .zip file on the Google Classroom.	

Problem # 1


86.4%
86.4% OF PEOPLE WILL BELIEVE ANY DATA YOU PUT IN A POWERPOINT SLIDE, EVEN IF YOU JUST TOTALLY MADE IT UP TO PROVE YOUR POINT.
TOM FISHBURNE

© marketoonist.com

Download "PakWheels Automobile" dataset from Kaggle: (https://www.kaggle.com/datasets/muhammadwaqargul/pakwheels-used-car-dataset-october-2022) and perform various kinds of inferential statistics tasks stated below to demonstrate your knowledge and understanding.

a. **Cleaning, descriptive statistics and exploratory analysis:** Perform data cleaning if required and provide a basic data understanding using descriptive statistics and visualization as follows.

   i. Visualize categorical attributes using the most appropriate charts.

ii. Provide summary statistics for numeric attributes (mean, media, SD) and visualize distributions (histograms, boxplots).

iii. Comment on skewness, outliers, and whether data transformations might be needed (but do not perform any transformation).

b. **One-sample inference:** Pick any attribute of your interest, a suitable test statistic and a hypothesized (but meaningful) value for the population. Perform one-sample inference (z-test/t-test, whichever is more appropriate). Include a 95% confidence interval for the test statistic and interpret the results.

c. **Two-sample comparison:** Pick two features and perform a two-sample comparison (z-test/t-test). Formulate hypotheses, test using a two-sample test (z-test/t-test, whichever is more appropriate). Test your null hypothesis with a 5% significance level using:

i. Critical value

ii. p-value

For both methods above, state your hypothesis results.

d. **Multiple-group comparison:** Using the ‘title’ attribute* (e.g., Toyota, Honda, Suzuki) or ‘model_year’ (e.g., ≤2010, 2011-2015, ≥2016; build suitable bins), test whether ‘price’ differs across these groups. Check assumptions (normality, homogeneity of variances). If violated, use a Kruskal-Wallis test. Formulate and test hypotheses with 5% and 1% significance levels, and interpret the results.

* You might need to perform data cleaning for this feature to extract model/brand name.

e. **Effect of significance level:** From Part (d), analyze and report if changing the significance level had any impact on your hypothesis results or not.

f. **Correlations:** Plot scatter chart for all pair of numeric attributes. Compute correlations between all the numeric features and show the results in a heatmap chart. Report correlation coefficients, p-values, and CIs and interpret the results.

g. **Bootstrapping:** Repeat any of the hypothesis test performed above using bootstrapping (e.g., 1000 iterations), make your decision based on your analyses of data distributions in Part (a).

Compare your results with the original test perform above and state your analysis if and why bootstrapping has any effect.

Problem # 2

Correlation might be misleading, and you will explore and test about them in this problem.

Spurious Correlations
SALES
SHAVED HEADS
WE FOUND THIS CORRELATION IN THE DATA. EVERYONE TAKE A RAZOR.

Source: http://bit.ly/3Yjy3EX

a. Explore **Spurious Correlations** (https://www.tylervigen.com/spurious-correlations), identify and report one of the correlations which in your opinion if were true would have an impact in the world. Your response should contain snapshot of the original chart and 1-paragraph description justifying your response.

b. State your learning after exploring spurious correlations. How likely we can expect to see such correlations in a real dataset? And how can such “wrong” correlations be identified?

c. Explore real datasets on Kaggle and create a new spurious correlation with r >= 0.8. Provide a link to the datasets, attributes used, #of records and draw scatter chart.

d. Create a 95% confidence level of the correlation and compute significance of the correlation using p-value.

e. Perform bootstrapping for 1000 iterations to estimate the 95% confidence interval for the correlation coefficient between the two variables using percentiles. State your findings in relation to the observed r.