EGN 5458: STATISTICAL APPLICATIONS FOR ENGINEERS

MIDTERM 2
Due: 11.23.2023, 08:00 PM

Please submit your report on CANVAS using the assignment page for Midterm 2. Please submit a SINGLE Word or PDF file that includes your MATLAB command window screenshots and plots showing your RUN RESULTS as well as your SCRIPTS.

PLEASE ALSO NOTE THAT YOUR MIDTERM SHOULD INCLUDE VISIBLE SCANS AND CLEAR SCREENSHOTS.

You should not calculate the descriptive characteristics of a data set using the command window. Please write a MATLAB script that calculates these parameters for given data and change that input data set for different problems.

Please EMBED YOUR SCRIPT (CODE) into the text with your submission instead of a separate attachment.

1. **(65 pts) THIS QUESTION IS LINKED TO MIDTERM 1 FIRST QUESTION! (TRUCK-INVOLVED CRASH STUDY)** As part of Midterm 1, we conducted a comprehensive analysis of the complex dynamics behind truck-involved crashes, recognizing their severe implications. This analysis examined a dataset derived from Florida roadway network, which contains 5-year truck-involved crashes along with diverse explanatory variables, including characteristics of rest areas, travel attributes, and distances from crash sites to rest areas. Your task in this question is to write a MATLAB script to investigate the contribution of related factors and explore the correlation between truck-related crashes and various influential factors. This study intends to address the crucial issue of driver sleepiness and its potential contribution to such crashes. The findings of the study could therefore assist safety officials in determining the most appropriate locations for future rest areas and truck parking on interstate highways. The dataset contains entries for each rest area, detailing the number of truck-related crashes in their vicinity. Students were provided with an individualized Excel sheet (refer to Midterm 1) for their analysis. Please refer to Midterm 1, “Variable Explanations” table for further information regarding the variables.

   File Name: Midterm1_Fall2023_EGN5458_Data_1.xlsx

   a. Run a correlation analysis to determine the correlation between certain variables (See below table) and **TR_CRASH_CT**s. Use the provided correlation command in MATLAB. Present the correlation values of the given variables in a table. The resulting table should resemble the example shown below. Ensure that your correlation table includes values representing the correlations between the selected variables and **TR_CRASH_CT** (excluding “Row_ID” from the correlation analysis).

```
R = corrcoef (A)
```

https://www.mathworks.com/help/matlab/ref/corrcoef.html

Variable Name	**Correlation Coefficients with TR_CRASH_CT**
AADT	**0.15**
AVG_SEV	**-0.12**
WEIGH_ST_BI	**-0.08**
SAFE_SEC	**0.001**
FOOD_BI	**…**
TOT_LOT	
DIST_URB	
AVG_DIST	

Note: Given values are example only, not the actual result.
Make sure the variables are sorted in descending order based on the absolute values of correlation coefficients.

b. The following formulas have been used in Midterm 1 to define the additional variables **TR_CRASH_RATE** and **REST_SAT**:

$$TR\_CRASH\_RATE_i = \frac{1{,}000{,}000 * TR\_CRASH\_CT_i}{365 * N * AADT_i}$$

$$REST\_SAT_i = \frac{(2.68 \times SAFE\_SEC_i) + (3.26 \times WEIGH\_ST\_BI_i) + (8.4 \times TOT\_LOT_i)}{1000}$$

Please refer to the specified variables and perform a correlation analysis to determine the correlation between selected variables (See below table) and the calculated **TR_CRASH_RATE**s. Use the provided correlation command in MATLAB. Present the correlation values of the given variables in a table. The resulting table should resemble the example shown below. Ensure that your correlation table includes values representing the correlations between all variables and **TR_CRASH_RATE** (excluding “Row_ID” from the correlation analysis).

Variable Name	**Correlation Coefficients with TR_CRASH_RATE**
AVG_DIST	**0.23**
AVG _SEV	**-0.19**
TOT_LOT	**-0.12**
SAFE_SEC	**0.003**
FOOD_BI	**…**
WEIGH_ST_BI	
DIST_URB	
REST_SAT	

Note: Given values are example only, not the actual result.
Make sure the variables are sorted in descending order based on the absolute values of correlation coefficients.

c. Note that correlation coefficients could range between -1 and 1 depending on the positive and negative correlation. “-1” and “1” show highest possible correlation, while 0 is the smallest. Discuss similarities and differences between the results obtained in Part (a) and those derived from Part (b). Highlight any variables that appear consistently influential or contradictory in their impact on truck-related crash occurrences.

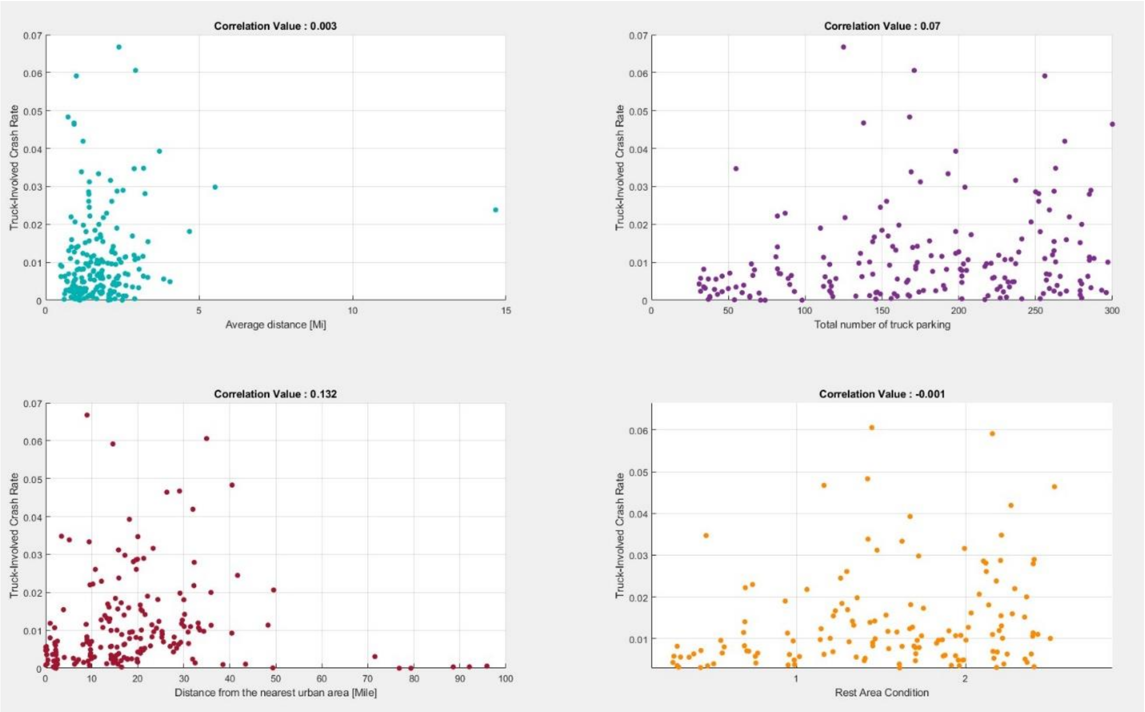
d. Compute the normalized values for **TR_CRASH_RATE** using the formula for Min-Max normalization (Ensure that the resulting values fall within the range of 0 to 1):

$$NORM\_TR\_CRASH\_RATE_i = \frac{TR\_CRASH\_RATE_i - MIN_{TR\_CRASH\_RATE}}{MAX_{TR\_CRASH\_RATE} - MIN_{TR\_CRASH\_RATE}}$$

e. Using a boxplot to vissually illustrate the spread, central tendency, and any potential outliers in the **NORM_TR_CRASH_RATE** values. Utilize the boxplot function to generate a graphical representation of the distribution of the NORM_TR_CRASH_RATE variable. Disscuss the result.

f. **<u>(DEAL WITH OUTLIERS!)</u>** Calculate the **<u>95th percentile</u>** of the calculated **NORM_TR_CRASH_RATE** values. The 95th percentile represents a value below which 95% of the data falls. Then create a subset of the dataset by eliminating all rows where the associated **NORM_TR_CRASH_RATE** values are greater than the computed 95th percentile value. Ensure that the new subset contains only the rows meeting the specified criterion. Breifly explain the significance of this process.

g. Utilizing the subset created in Part (f), proceed to employ the scatter command to generate separate plots representing the correlation between the **TR_CRASH_RATE** and the four most highly correlated variables, as determined by the correlation coefficients calculated in Part (b). Please produce four distinct plots, each illustrating a numerical variable against the **TR_CRASH_RATE**. Employ the subplot command to incorporate these four plots into a single figure. Ensure distinct colors and markers for each plot, distinguishing between the representations. Refer to the example plot provided below for guidance. Label the x- and y-axes appropriately, designating **TR_CRASH_RATE** on the y-axis..



Note: Your figures do not need to be identical, but they need to have similar characteristics.

h. Consider the new subset created in Part (f) and use Curve Fitting Toolbox (Command in MATLAB: cftool) to create polynomials for the relationship between **TR_CRASH_RATE** and four most correlated variables selected in Part (g) (**TR_CRASH_RATE** on y-axis) with a degree of 1, 2, and 3. DISCUSS your findings.

i. Based on the abovementioned numerical variables and corresponding **TR_CRASH_RATE**, INTERPRET your findings and COMMENT on the results. What can be deducted from results obtained in Part (g) and Part (h)? What policies or plans can be implemented based on your findings in order to prevent truck-involved crashes in the vicinity of rest area? Please provide your comments for each variable individually in addition to a general discussion on your findings.

2. (35 pts) **THIS QUESTION IS LINKED TO MIDTERM 1 LAST QUESTION!** Car manufacturing companies often work on real-life data to be able to come up with a model and effective factors to determine the best mile-per-gallon design for a car. Keeping that purpose in mind, load **carbig** data available in the MATLAB database. It includes measurements for cars between years 1970 and 1982. You will have six data sets: acceleration, cylinders, displacement, horsepower and weight as the predictor variables and mpg as the response variable (Take confidence level as 95% for all the steps).

a. For all numerical data sets, conduct hypothesis testing on the mean and standard deviation, and comment on the results. You can use the mean and standard deviation of the fitted normal distribution functions (from the previous midterm using dfittool) in order to create your hypotheses.

b. You used cftool to plot one predictor variable versus the response variable one by one and show the resulting plots on the same figure (acceleration vs. MPG, cylinders vs. MPG, etc.). Now, fit functions on the predictor vs. response variable plots (linear, different polynomials, normal, and exponential). Provide the necessary goodness of fit information (SSE, MSE, RMSE, R^2, and adjusted R^2). Comment on the results.

c. Plot time series of all the numerical predictor variables and response variable in the same figure using subplots (Variable versus time (year)). Comment on the changes with respect to time.

d. Now, load **carsmall** data which only includes three years data for the variables: 1970, 1976 and 1982. Consider this as a sample taken from **carbig** data. Perform the same steps starting from (a) to (c) on the carsmall data.

e. Discuss the differences between the carsmall data and carbig data. Suppose that **carbig** is your population and **carsmall** is your sample. Discuss the results. Are there any difference in hypothesis testing, fitting and estimation analysis results? Do you think the sample data (**carsmall**) accurately represents the whole data (**carbig**)?