STA234 HW-3

DUE DATE: 2/27/2025 by 10PM in Moodle

Problem 1 [5 points]

During feature (column) selection using the following dataframe (named sample), “Column1” and “Column2” proved to be non-significant. Hence, we would not like to take these two features into our predictive model. **Show in R how will you select all the rows from column 3 to column 6 for the below dataframe named table?**

sample

	Column1	**Column2**	**Column3**	**Column4**	**Column5**	**Column6**
Name1	Alpha	12	24	54	0	Alpha
Name2	Beta	16	32	51	1	Beta
Name3	Alpha	52	104	32	0	Gamma
Name4	Beta	36	72	84	1	Delta
Name5	Beta	45	90	32	0	Phi
Name6	Alpha	12	24	12	0	Zeta
Name7	Beta	32	64	64	1	Sigma
Name8	Alpha	42	84	54	0	Mu
Name9	Alpha	56	112	31	1	Eta

Problem 2 [30 points]

We will use the PIMA dataset which consists of a population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. There are nine variables, namely

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function

8. Age (years)

9. Class variable for diabetic or not according to WHO (0 or 1)

Starting with the work you did in HW-2,

(a) Import the data from Moodle or shared Google drive, it is called pima.csv. Change the name of the nine columns to preg_times, glucose_test, blood_press, tsk_thickness, serum, bm_index, pedigree_fun, age, class.

(b) All patients (768 Observations) in this dataset contains are females at least 21 years old of Pima Indian heritage. All zero values for the biological variables other than number of times pregnant should be treated as missing values. Count how many zeros are there in each variable (column). For any 0 in the data (except for class and preg_times) assign it as an NA.

(c) For class variable, check if it is a factor and if not, then make it a factor with levels 0 replaced with neg (for negative diabetic) and 1 replicated with pos (for positive diabetic),

(d) Make data subsets for four age groups: 21-36, 37-51, 52-66 and 67-81.

(e) Create a new factor vector called age.factor, with age in pima data replaced with the age group.

Now do the following in R:

(f) **5 points** Using the age.factor variable in ggplot, make a barplot for four age groups: 21-36, 37-51, 52-66 and 67-81 indicating the number of women in each age group.

(g) **5 points** Make a histogram of BMI for all women using ggplot function with percentage on the y-axis.

(h) **5 points** Make a histogram for the BMI of women with different color for each age group with percentage on the y-axis.

(i) **5 points** Make comparative boxplots for blood pressure of women in four age groups.

(j) **5 points** Make a scatterplot between blood pressure (y) and BMI (x) using separate colors for different age groups. Comment on the relation.

(k) **5 points** Make a layered scatterplot between blood pressure (y) and BMI (x) using separate colors for different age groups and add fitted regression least squares line. Comment of the relations.

Problem 3 [30 points]

Using the RailTrail dataset from mosaicData package. You need to install the package using install.packages(“mosaicData”) in your Rstudio Console, before you run the functions below:

```
library(mosaicData)
head(RailTrail)
```

```
##   hightemp lowtemp avgtemp spring summer fall cloudcover precip volume wee
kday
## 1       83      50    66.5      0      1    0        7.6   0.00    501    
TRUE
## 2       73      49    61.0      0      1    0        6.3   0.29    419    
TRUE
## 3       74      52    63.0      1      0    0        7.5   0.32    397    
TRUE
## 4       95      61    78.0      0      1    0        2.6   0.00    385   F
ALSE
## 5       44      52    48.0      1      0    0       10.0   0.14    200    
TRUE
## 6       69      54    61.5      1      0    0        6.6   0.02    375    
TRUE
##   dayType
## 1 weekday
## 2 weekday
## 3 weekday
## 4 weekend
## 5 weekday
## 6 weekday
```

1) Create a scatterplot of the number of crossings per day volume against the high temperature that day. Please note that you can use ?RailTrail to find out more about the dataset.

2) Separate the plot into facets by weekday.

3) Add least square fitted regression lines to the two facets.

4) Summarize the information that the data graphic from question 3 conveys.

Problem 4 [15 points]

The MLB_teams dataset in the mdsr package contains information about Major League Baseball teams in the past four seasons. There are several quantitative and a few categorical variables present. You may need to install the package using install.packages("mdsr") in your Rstudio Console, before you run the functions below:(Please note that you can use ?MLB_teams to find out more about the dataset.)

```
library(mdsr)
head(MLB_teams,4)

## # A tibble: 4 x 11
##   yearID teamID lgID      W     L  WPct attendance normAttend payroll metr
oPop
##    <int> <chr>  <fct> <int> <int> <dbl>      <int>      <dbl>   <int>    <
dbl>
## 1   2008 ARI    NL       82    80 0.506    2509924      0.584  6.62e7   448
9109
```

```
## 2   2008 ATL   NL       72    90 0.444    2532834      0.589  1.02e8  561
4323
## 3   2008 BAL   AL       68    93 0.422    1950075      0.454  6.72e7  278
5874
## 4   2008 BOS   AL       95    67 0.586    3048250      0.709  1.33e8  473
2161
## # ... with 1 more variable: name <chr>
```

```
names(MLB_teams)
```

```
##  [1] "yearID"     "teamID"     "lgID"       "W"          "L"
##  [6] "WPct"       "attendance" "normAttend" "payroll"    "metroPop"
## [11] "name"
```

1) Make a scatterplot to illustrate the relationship between winning percentage and payroll in context.

2) Add the league in which team played to show more information to make layered or facets. Add smoothed regression line to show the trends.

Problem 5: [20 points]

Using the mpg dataset in ggplot2 package answer the following:

```
library(ggplot2)
data(mpg)
```

(a) Do cars with big engines use more fuel than cars with small engines? Create a scatterplot in ggplot to justify your answer.

(b) To display the class of each car, use colors in the above scatterplot of displ versus hwy variables.

(c) Use facets to display the scatter plots for the class of each car.

(d) Using geom_smooth() to make scatter plot for displ vs hwy for each category in variable drv which describes a car's drivetrain. Use default method (do not specify method=lm) to get curved fits. Check class of drv variable and make sure it is a factor so R can make the right plot for all levels.

Project Problem: [30 points]

For your data project,

1. Introduction: Tell me what problem you are working on? Why is this problem interesting and important. State specific research questions your group will work on. Introduce recent research done in area related to your problem. You can pack all this together to motivate us. Do keep it short, to the point, and interesting.
2. Data: Tell me about the data resource and explain dimensions of the data, variables in the data, and how does this data relate to your research questions.
3. EDA: Use your dataset to make data visualizations that explain the variables of interest and how information through the graphics provides easy solution for your research questions. Explain your steps on how these visualizations help with your project.