Test Paper Set 5

25 Marks

1. Load the data

2. create an analysis to show which gender has highest customers from the bank

3. create a Age analysis using Histogram to undertand the distribution of customers

4, create a Balance analysis using Histogram to undertand the distribution of customers with different balance

5. Create a chart to regional analysis

25 Marks

Instructions

Just complete the tasks outlined below.

Get the Data

You can download the data or just use the supplied csv in the repository. The data has the following features:

- datetime - hourly date + timestamp
- season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
- holiday - whether the day is considered a holiday
- workingday - whether the day is neither a weekend nor holiday
- weather -
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp - temperature in Celsius

- atemp - "feels like" temperature in Celsius
- humidity - relative humidity
- windspeed - wind speed
- casual - number of non-registered user rentals initiated
- registered - number of registered user rentals initiated
- count - number of total rentals

Q1. Read in bikeshare.csv file and set it to a dataframe called bike.

Q2. Check the head of df

Q3. Create a scatter plot of count vs temp. Set a good alpha value.

Q4. Plot count versus datetime as a scatterplot with a color gradient based on temperature. You'll need to convert the datetime column into POSIXct before plotting

Q5. What is the correlation between temp and count?

Q6. Create a boxplot, with the y axis indicating count and the x axis begin a box for each season.

Q7. Create an "hour" column that takes the hour from the datetime column. You'll probably need to apply some function to the entire datetime column and reassign it. Hint:

```
time.stamp <- bike$datetime[4]

format(time.stamp, "%H")
```

Q8. Create the same plot for non working days.

Q9. Use lm() to build a model that predicts count based solely on the temp feature, name it temp.model

Q10. Get the summary of the temp.model

Q11. Finally build a model that attempts to predict count based off of the following features. Figure out if theres a way to not have to pass/write all these variables into the lm() function