MA 380 Project Two

Bike Sharing Analysis

Your Name Must be Here

Due Date: Friday November 17, 2023, 11:59 PM EST

Purpose

This assignment has 10 tasks (numbered 0 to 9) worth a total of 200 points. Each task shows the number of points attached to it. There are there key activities for the assignment:

1. comprehensive analysis of a data set
2. interpretation of key findings
3. communication of analysis results to stakeholders

During your career you will be involved in all three activities and this assignment's purpose is to introduce you and give the tools and experience to successfully navigate them.

Skills

This assignment will help you practice the following skills that are essential to your success in this course and in your professional life beyond school:

1. Apply exploratory tools to determine which predictor variables may help us determine the target variable
2. Understand how different predictor variables are related to each other and how they affect the prediction of the target (aka response) variable
3. Use statistical software to fit a model to the data
4. Employ diagnostic tools to improve a statistical model
5. Interpret the parameters of a statistical model and communicate your findings
6. Use RStudio and R Notebook to author a document that integrates code and prose into single text file. This approach goes a long way to ensuring the reproducibility of your findings.

Knowledge

This assignment will also help you to become familiar with the following important content knowledge in this discipline:

1. Count models within the Generalized Linear Modeling Framework
2. Apply feature engineering techniques to predictor variables to extract more information
3. Use various residuals to evaluate and improve a statistical model

Criteria for Success

A successful project will produce an R Notebook that is clear, concise, and informative. Each task will be graded on the quality of your thought process, conclusions, and its presentation as documented in your written response. Answers to individual tasks can use technical language, but the answer to the last task,

where you need to summarize your your findings for your stakeholders, should be focused for a business audience and thus not use technical language.

It is very easy to produce lots of computer output such as tables or graphs. Following this voluminous strategy is a recipe for failure. You should **look at everything** but include the most relevant pieces and explain why some other items may not be relevant. Also producing output and not explaining it in plain English will diminish the impact of your work.

Each spelling error or misused word, at best says you are careless, and at worst diminishes the credibility of your work. After a few such errors, your readers may give up and and move on to other activities. I will skip over the first three spelling mistakes (or misused words), but thereafter, each misspelling or misused word will deduct $10(K-3)$ points from your overall score, where K is the number of such errors.

Do not worry about about formatting issues, such as page breaks, heading sizes, and similar items. But do worry about the graphs you are producing for this report. Make sure that your labels are correct and informative.

Recommendations are stated in unambiguous language. The writing style should be direct and it should not narrate the steps you are taking. Rather you should focus on conveying the main facts and selections used.

What to Submit

You should only upload your R Notebook into Brightspace. Please use the following naming convention for your file: `ma380-pr02-<lst name>-<fst name>.Rmd` where `<lst name>` represents the first **three** letters of your last name and `<fst name>` corresponds to the first **two** letters of your first name. For example, my submission file should be named: `ma380-pr02-sch-er.Rmd`.

I will knit your submission into a Word or PDF file and provide feedback on that document. It is important that your R Notebook is error free. If I cannot knit your submission, then I will return it to you to fix and deduct 10 points from your overall score. I highly recommend that you knit your document to either HTML or Word and take a close look at the final rendered document to make sure that is what you want to submit.

Business Problem

Many cities and towns now provide locked bikes throughout their neighborhoods. Customers sign-up for a sharing contract and they are able to pick up a bike in one location, and ride it to a different location to return it. You have been hired to help a town understand when are customers using the bikes. The town administration would like to create a model that will predict the number of bikes used in a given hour and the locations that the bikes are moving from and to.

The town's Information Technology department has prepared a data set for you. The data dictionary is provided below. This is the only data you have available for your analysis.

Data Dictionary

Variable	Description
season.code	Season (1 = Winter, 2 = Spring, 3 = Summer, 4 = Fall
year.code	Year indicator (0 = 2011, 1 = 2012)
hour	Hour (integer 0 to 23)
holiday.code	Indicator of holiday (0 = No, 1 = Yes)
weekday.code	Day of the week (0 = Sunday, 1 = Monday, . . . , 6 = Saturday)
weathersit.code	Weather situation (1 = Clear/Partly Cloudy, 2 = Mist, 3 = Rain or Snow)
temp	Normalized temperature in Celsius. [(t - t_min)/(t_max - t_min), t_min = -9, t_max = 39]
humidity	Normalized humidity. Values are divided by 100 (max possible)

Variable	Description
windspeed	Normalized wind speed. Values are divided by 67 (max possible)
bikes	Count of rental bikes in each hour

Task 0 (0 points)

Read the data and provide appropriate types to the variables in the data set. Many of the variables are code as integers making it difficult to know what their values mean. Create new variables that are more human friendly. Note that in later tasks you may need to modify the variable types given here. Use function `read_csv()` from the `tidyverse` package and set the argument `col_types` appropriately.

Task 1 (10 points)

Assess whether or not the data you have will help you address the business problem that the town is facing. In your assessment be sure to clearly mention how the data you have been given will be useful or not in addressing the two concerns that the town's administration has.

Task 2 (10 points)

Which variables should be treated as categorical? Provide an explanation for choosing these variables as categorical and change their types in your data set.

Task 3 (12 points)

Create a new variable called `workday` with values of `Yes` if the day is indeed a workday and `No` if it is either a weekend or a holiday. Describe one advantage and one disadvantage in including `workday` in your model.

Task 4 (30 points)

Conduct an exploratory data analysis on the information you have available with a focus on answering some of the key questions that the town's administration has. Select **three** graphs and, for each one of them, explain what modeling decisions it supports.

Task 5 (24 points)

Explore the **mean-variance** relationship for the number of bikes rented per hour. Provide a bivariate plot showing this relationship. For each of the Poisson, Negative Binomial, and Gamma distributions use the information in the mean-variance relationship to determine which of these distributions would be most suitable for building a generalized linear model.

Task 6 (40 points)

Based on your responses to the previous tasks select an initial model (write it down here) and then search for a good model of the number of bikes rented each hour. Select your final model and perform a thorough diagnostic analysis.

Task 7 (16 points)

For a general audience interpret your final model from the previous task.

Task 8 (8 points)

Some variables were not included in your final model. Select two of them and explain why you did not include them. Back up your argument with either a table or a graph.

Task 9 (50 points)

Write a short summary of your findings that you would share with the town administrators. Be sure to address a general audience and to focus your recommendations on solving the business problem they face.

Your written comments should not exceed 750 words. You may include two graphs and/or tables to support your arguments.
