STA130H1S – Fall 2022

Problem Set 7

() and STA130 Professors

Instructions

Complete the exercises in this .Rmd file and submit your .Rmd and .pdf output through Quercus on Thursday, November 3rd by 5:00 p.m. ET.

```
library(tidyverse)
```

Part 1: Broadway

Question 1: Driving on the "Right" side of the Road

Lin-Manuel Miranda was nominated for "Best Original Song" for the March 27, 2022 the Academy Awards (also known as the Oscars) for his work on the Disney movie Encanto. Miranda had already won an Emmy, Grammy, and Tony (mostly for his work on the broadway musical "Hamilton"), so he was very close to the (EGOT)[https://www.vanityfair.com/hollywood/2022/02/oscar-nominations-2022-will-lin-manuel-miranda-finally-egot-for-encanto] (Emmy, Grammy, Oscar and Tony), a rare occurrence as only 16 people have won all four awards see here. Unfortunately, Miranda did not win the Oscar in 2022. Perhaps he will soon!

In this question we will look at a sample of weekly broadway musical data available in the `broadway.csv`. This data set contains a sample of Broadway musical information for 500 weeks from 1985 to 2020. In this data set an observation is one broadway musical in a particular week (ending on a Sunday). Variables of interest are:

- show: Name of the broadway musical/show.
- Hamilton: indicates whether the musical is "Hamilton" or not.
- week_ending: Date of the end of the weekly measurement period. Always a Sunday.
- weekly_gross_overall: Weekly box office gross for all shows.
- avg_ticket_price: Average price of tickets sold in a particular week.
- top_ticket_price: Highest price of tickets sold in a particular week.
- seats_sold: Total seats sold for all performances and previews in a particular week.
- pct_capacity: Percent of theatre capacity sold. Shows can exceed 100% capacity by selling standing room tickets.

Let's explore different ways to estimate the average ticket price for Broadway shows!

```
broadway_data <- read_csv("broadway.csv")
glimpse(broadway_data)
```

```
## Rows: 500
## Columns: 8
## $ show                  <chr> "La Cage aux Folles", "42nd Street", "42nd Street~
## $ Hamilton              <chr> "No", "No", "No", "No", "No", "No", "No", "No", "~
## $ week_ending           <date> 1985-07-28, 1985-09-08, 1985-09-15, 1985-12-15, ~
## $ weekly_gross_overall <dbl> 2989271, 2474396, 2844860, 4169643, 3555363, 3632~
## $ avg_ticket_price     <dbl> 34.54, 30.31, 30.50, 35.00, 27.74, 16.60, 17.19, ~
## $ top_ticket_price     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ seats_sold           <dbl> 11841, 7251, 7890, 10846, 2803, 2204, 5740, 10861~
## $ pct_capacity         <dbl> 0.8795, 0.5477, 0.5959, 0.8056, 0.2967, 0.4364, 0~
```

(a) Make a plot showing the relationship between the average ticket price (on the y-axis) and the weekly gross overall sales (on the x-axis).

```
# code you answer here
```

(b) Briefly explain whether or not you think it is appropriate to characterize and summarize the association in the above plot with a straight line.

REPLACE THIS TEXT WITH YOUR ANSWER

(c) Use the `mutate()` function to add the new variables `log_avg_ticket_price = log(weekly_gross_overall)` and `weekly_gross_overall_in_hundred_thousands=weekly_gross_overall/100000` to the dataset.

```
# code you answer here
```

(d) Plot the association between `log_avg_ticket_price` (on the y-axis) and `weekly_gross_overall_in_hundred_th` (on the x-axis) and use `geom_smooth(method=lm, se=FALSE)` to add a line of best fit to the plot. Describe the association you observed in the plot.

Note: You will learn more about transforming variables in future courses and are not required to be able to explain why we've done this here. You can just treat `log_avg_ticket_price` as we have other variables in class and refer to it as "the natural log of average ticket price".

```
# code you answer here
```

REPLACE THIS TEXT WITH YOUR ANSWER

(e) Use the `cor()` function to calculate the correlation between `broadway_data$log_avg_ticket_price` and `broadway_data$weekly_gross_overall_in_hundred_thousands`.

```
# code you answer here
```

(f) Write down a simple linear regression model specification with response `log(avg_ticket_price)` and explanatory varaible `weekly_gross_overall_in_hundred_thousands`. Explain each compnwnt of the model.

REPLACE THIS TEXT WITH YOUR ANSWER

Hint: If you copy math equations from another software into your .Rmd document, you'll get errors when trying to knit. Instead, you should type your math equations directly in your .Rmd document. Here are some tips and examples for doing this:

1. In a .Rmd document, math equations and symbols must be typed between dollar symbols ($).
2. If you want your equation/symbol to appear in the middle of a sentence, use only one dollar sign before and one dollar sign after. For example, we can typeset beta-hat-0 in .Rmd as $\hat{\beta}_0$.
3. If you want your equation to appear on a line on its own, type it on a separate line and put two dollar signs at the begining and the end. For example,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1$$

4. A few other useful symbols you may need in this question are epsilon (ϵ), “not equal” ($\neq$), and superscripts (e.g. i^{th}).

(g) State the null and alternative hypotheses you would use assess whether the slope of the linear regression model where weekly gross overall income in 1000s is predicting the log average ticket price.

REPLACE THIS TEXT WITH YOUR ANSWER

(h) Use R to fit the linear model that corresponds with your line of best fit above. Report the fitted equation of the line. Interpret the regression coefficients in the context of this data AND make a conclusion about the hypotheses you defined above.

REPLACE THIS TEXT WITH YOUR ANSWER

```
# code you answer here
```

(i) Get the R^2 for your model and write one sentence interpreting it in context.

```
# code you answer here
```

(j) Create a plot of the association between `top_ticket_price` (on the y-axis) and `top_ticket_price` (on the x-axis) faceted by whether the musical was “Hamilton” or not using `facet_wrap(~Hamilton)`.

```
# code you answer here
```

(j) Create a plot of the association between `top_ticket_price` (on the y-axis) and `top_ticket_price` (on the x-axis) faceted by whether the musical was “Hamilton” or not using `facet_wrap(~Hamilton)`.

Calculate the correlation between top ticket price and average ticket price for both Hamilton and non-Hamilton musicals using the `group_by`, `summarise()` and `cor(top_ticket_price, top_ticket_price, na.rm=TRUE)`.

```
# code you answer here
```

Part 2: Optional Indicator Variable Simple Linear Regression

You may complete these questions for practice if you wish. ***You are not required to complete these questions as they ARE NOT included as part of your mark.***

Question 2: Starbucks

The `starbucks.csv` dataset contains data on calories and carbohydrates (in grams) in Starbucks food menu items.

```
starbucksdata<-read_csv("starbucks.csv")
glimpse(starbucksdata)
```

```
## Rows: 77
## Columns: 7
## $ item     <chr> "8-Grain Roll", "Apple Bran Muffin", "Apple Fritter", "Banana~
## $ calories <dbl> 350, 350, 420, 490, 130, 370, 460, 370, 310, 420, 380, 320, 3~
## $ fat      <dbl> 8, 9, 20, 19, 6, 14, 22, 14, 18, 25, 17, 12, 17, 21, 5, 18, 1~
## $ carb     <dbl> 67, 64, 59, 75, 17, 47, 61, 55, 32, 39, 51, 53, 34, 57, 52, 7~
## $ fiber    <dbl> 5, 7, 0, 4, 0, 5, 2, 0, 0, 0, 2, 3, 2, 2, 3, 3, 2, 3, 0, 2, 0~
## $ protein  <dbl> 10, 6, 5, 7, 0, 6, 7, 6, 5, 7, 4, 6, 5, 5, 12, 7, 8, 6, 0, 10~
## $ type     <chr> "bakery", "bakery", "bakery", "bakery", "bakery", "bakery", "~
```

(a) Produce a plot that shows the association between carbohydrates and calories in Starbucks menu items. Describe this association.

```
# code you answer here
```

(b) Before calculating anything, estimate the correlation coefficient between carbohydrates and calorie content in Starbucks menu items based on the plot you produced in (a). Justify your answer.

REPLACE THIS TEXT WITH YOUR ANSWER

(c) Calculate the correlation between carbohydrate and calorie content of Starbucks menu items. How does this compare to your estimate in part (b)?

REPLACE THIS TEXT WITH YOUR ANSWER

```
# code you answer here
```

(d) Write down a simple linear regression model specification for the content of Starbucks menu items with `calories` as tbe response variable and `carb` as the explanatory variable. Explain each term in the model.

REPLACE THIS TEXT WITH YOUR ANSWER

(e) Describe what the test $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$ is testing.

REPLACE THIS TEXT WITH YOUR ANSWER

(f) Use R to fit the regression model in (d) to these data. Report the fitted regression line. Interpret the regression coefficients in the context of this study AND make a conclusion about the hypotheses you defined above.

REPLACE THIS TEXT WITH YOUR ANSWER

```
# code you answer here
```

(g) Add the estimated linear regression line that you calculated in (f) to the plot you generated in (a). Compute the coefficient of determination, R^2. How well does the linear regression line seem to capture the relationship between `carb` and `calories`? Justify your answer.

REPLACE THIS TEXT WITH YOUR ANSWER

```
# code you answer here
```

(h) Based on the Starbucks data, create a new dataset called `starbucks_lunch` which only contains food items which are only of the "sandwich" or "bistro box" type. Create a boxplot comparing the distribution of calories for these two types of items.

```
# code you answer here
```

(i) Fit a linear regression model to test whether there is a difference in mean calories for items of type "bistro box" and items of type "sandwich". Write a sentence summarizing your conclusion.

REPLACE THIS TEXT WITH YOUR ANSWER

```
# code you answer here
```