**Analytics Test – Nov 2021**

PLEASE READ THIS PAGE FIRST

**Instructions:**

Go through the questions as quickly as you can without sacrificing accuracy. The majority of candidates will not complete all questions correctly, and not finishing all questions does not mean you failed the test. The questions vary in difficulty, so if you are struggling with a question, you might want to consider moving on to the next one and returning to it if you finish with time to spare. Do not spend too much time formatting/commenting your code, and instead focus on completing the given questions.

**Data:**

All the files you require will be in the Zipped folder which you will be able to download from the provided link during the test.

You have been provided with the file **sales\_labour\_costs.csv** containing weekly sales and labour costs, and product count by department for grocery stores of a US retailer. The file **store\_characteristics.csv** contains some additional information regarding the stores.

The tables follow a structure of a mostly normalised relational database of the form we often encounter in client situation. A data dictionary is provided in the appendix.

**Question Block A: Data cleaning & exploration**

1. Please load the 2 given files into R / Python (sales\_labour\_cost.csv, store\_characteristics.csv). How many columns does the sales\_labour\_cost.csv data have?
2. How many **more** store id’s are there in the sales\_labour\_cost.csv compared to the store\_characteristics.csv?
3. What is the oldest date in sales\_labour\_cost.csv?
4. This question involves cleansing the State column in store\_characteristics.csv
5. Using store\_characteristics.csv, how many stores don’t have any State information in store\_characteristics.csv? (i.e. the State value is NA)
6. How many stores does the State with the most stores have? (Hint: the capitalisation of the State values isn’t consistent)
7. For stores that have no State, replace the State with the most common State value given in a). What is the average Selling Area for stores in the most common State (including the stores whose values have been updated to this State)?
8. How many rows in sales\_labour\_cost.csv have either zero sales or zero labour costs?
9. This question involves cleansing the Department column in sales\_labour\_cost.csv
10. How many departments in sales.csv have no or negative sales AND no or negative labour costs? (aggregating across all weeks and stores)
11. Filter out these departments from the sales dataset (all rows of data for the stores, not just the weeks that don’t have any sales). What is the average weekly department sales value in this filtered sales dataset?
12. Run the code below to load the cleansed store characteristics and sales and labour cost files. (update the file path if needed). Group sales\_labour\_cost\_cleansed so that the data is at store-week level, rather than store-department-week level. Combine the sales and labour costs and store characteristics datasets, so that the rows from sales and labour costs are appended with new information from store characteristics dataset.

**R:**

library(readr)

store\_characteristics\_cleansed <- read\_csv("store\_characteristics\_cleansed.csv")

sales\_labour\_cost\_cleansed <- read\_csv("sales\_labour\_cost\_cleansed.csv")

**Python:**

import pandas as pd

store\_characteristics\_cleansed = pd.read\_csv("store\_characteristics\_cleansed.csv")

sales\_labour\_cost\_cleansed = pd.read\_csv("sales\_labour\_cost\_cleansed.csv")

1. Excluding date variables (Fiscal\_Year, Fiscal\_Week, Week\_Beginning) and stores with no product number data (where Number\_of\_Products is 0), which variable is most correlated with labour cost? (Hint: You will need to merge the sales and labour costs data with the store characteristics data to calculate this)
2. Create two new variables: 1) Sales per Product (using Sales and Number\_of\_Products), and 2) Labour Cost per Product (using Labour\_Cost and Number\_of\_Products).

Which variable is most correlated with Labour Cost per Product (excluding the original labour cost variable and the date-related variables)?

**Question Block B: Data Modelling**

1. Run the given code below to load the merged data and generate a linear regression model where the response variable is the weekly labour cost per product of a specific department within a store:

**R:**

model\_data1 <- read\_csv("model\_data1.csv")

lm\_bad <- lm(labour\_cost\_per\_product ~., data= model\_data1)

**Python:**

import statsmodels.api as sm

import(pickle)

X1 = pd.read\_pickle(“X1.pickle”)

Y1 = pd.read\_pickle(“Y1.pickle”)

lm\_bad = sm.OLS(Y1, X1)

fit\_bad = lm\_bad.fit()

* + 1. What is the unadjusted R squared of this model?
    2. Look at the structure of the data – the model is poor because considering the real- world meaning of the variables, not all are in the most useful data format (consider the difference between categorical and continuous variables). Correct this issue and re-run the model. What is the unadjusted R squared now? (Do not make any other changes like removing variables etc)

1. Run the given code below to load the merged data grouped by department and generate a linear regression model where the response variable is the weekly labour cost per product within a store:

**R:**

model\_data2 <- read\_csv("model\_data2.csv")

lm\_bad2 <- lm(labour\_cost\_per\_product ~., data=model\_data2)

**Python:**

X2 = pd.read\_pickle(“X2.pickle”)

Y2 = pd.read\_pickle(“Y2.pickle”)

lm\_bad2 = sm.OLS(Y2, X2)

fit\_bad2 = lm\_bad2.fit()

a) Considering what you have learned about this data so far, why might we want to remove the independent variable Sales from the regression?

* + - * 1. b) Suppose we wanted to remove any numeric variables which are not significant (pval >0.1). How many variables would this remove?

1. Run the code below to load a clean version of the model (*for all questions below use this model*):

**R:**

load("lm\_clean.rda")

**Python:**

lm\_clean = pickle.load(open(“lm\_clean.sav”,”rb”))

By interpreting this model:

* 1. What impact on labour cost per product is associated with a unit change in the number of products? (for a single store with all other variables remaining constant) (answer to 2 sig.fig.)
  2. On average, how much higher are labour costs per product expected to be in Florida (“FL”) compared to Mississippi (“MS”)? (for a single store with all other variables remaining constant) (answer to 2 sig.fig.)

1. Load the prediction table with the code below. Use the clean model from question 10 for this question.

**R:**

data\_pred <- read\_csv("data\_pred.csv")

**Python:**

data\_pred = pd.read\_pickle("data\_pred.pickle")

* 1. What does the model predict the labour cost per product to be for a store in Florida (FL) under Banner A with a selling area of 40,660 sqft, 2852 products, and $21.44 sales per product?
  2. What does the model predict to be the labour cost per product to be in the same scenario as above (q10a) but with 10% fewer products?

1. The client is interested in knowing what the labour cost per product would be if all variables apart from the number of products were held constant. What is the average predicted labour cost per product across the prediction table if all other predictors are held constant? (Hint: you should set the predictors to their average value, or their most common value if the predictor is categorical).

**Bonus Question:** How would you improve the model built in Question Block B?

**Appendix: Data Dictionary**

***Table Name:*** sales\_labour\_cost.csv

***Contains:*** Weekly sales and labour cost data, and product number data by department by store

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Store\_Number | Store ID |
| Department | Department ID |
| Fiscal\_Year | Financial year identifier to which the data is attributed |
| Fiscal\_Week | Financial week number to which the data is attributed |
| Week\_Beginning | Date of the starting day of the week to which the data is attributed |
| Sales | Total sales for that week in US Dollars |
| Labour\_Cost | Total labour cost for that week in US Dollars |
| Number\_of\_Products | Number of unique products stocked in the store’s department |
| Type | Anonymous tag related the department layout |
| Reset | Anonymous tag related to when the department was last reconfigured and reset |

***Table Name:*** store\_characteristics.csv

***Contains:*** Additional data related to the store and its region

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Store\_Number | Grocery Store ID |
| State | US State that the Store is located in |
| Banner | The branding/name of the store |
| Selling\_Area\_Sqft | Size of the store in Sq.ft. available for selling products |