**Analytics Test – Nov 2022**

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 the questions, 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. All datasets refer to properties that were sold between 2006 and 2010 in the US.

Files Required in Question Block A

* Property.csv – includes various details about the properties, such as Lot Area, Number of Rooms etc
* Garage.csv – includes details about the properties’ garages
* Transaction.csv – includes details about the sale of the properties, such as Year of Sale, Sales Price, etc.
* Combined.csv –a subset of the Property.csv and the Garage.csv datasets joined together

Files Required in Question Block B

* Model\_Cleansed.csv - a subset of the Property.csv, Garage.csv and Transaction.csv datasets joined together [for R only]
* Model\_Simple\_Cleansed.csv - same as above but with fewer variables [for R only]
* Model\_Test.csv – testing data with same columns as the above but excludes Sale Price [for R only]
* lm\_clean.rda – the linear regression model that was run using Model\_Simple\_Cleansed.csv [for R only]
* X1.csv – SalePrice column from the Model\_Cleansed.csv [for Python only]
* X2.csv – SalePrice column from the Model\_Cleansed.csv [for Python only]
* X2\_Test – testing data with same columns as Model\_Cleansed.csv [for Python only]
* Y1.csv – version of Model\_Cleansed.csv [for Python only]
* Y2.csv – version of Model\_Simple\_Cleansed.csv [for Python only]

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 3 given files into R / Python (Garage.csv, Property.csv, Transaction.csv). How many columns does the Property.csv data have?
2. What is the largest lot area in Property.csv?
3. How many **more** property Id’s are there in the Transaction.csv compared to the Garage.csv? (Hint: only count distinct values)
4. This question involves cleansing the Garage Type column in Garage.csv
   1. Using Garage.csv, what is the average Garage Area for the most common Garage Type? (Hint: the Garage Type spelling isn’t consistent)
   2. How many properties don’t have any Garage Type information in Garage.csv? (i.e. Garage Type value is either missing or NA)
   3. For properties that have no recorded Garage Type, assign the most common value. Once you have done this, what is the mean Garage Area for the most common Garage Type (including the newly assigned values).
5. How many properties in Property.csv have either zero Basement Full Bathrooms or zero Fireplaces?
6. What is the highest Sale Price of centrally airconditioned properties (hint: for these next two questions, you will need to join the datasets, and be wary of duplicate Ids)
7. How much higher is the average Property Price for properties built in 2007 vs 1900?
8. Run the code below to load the cleansed and merged property and garage datasets. (update the file path if needed).

**R:**

library(readr)

property\_combined <- read\_csv("Combined.csv")

**Python:**

import pandas as pd

property\_combined = pd.read\_csv("Combined.csv")

* 1. Which variable is the most correlated with Lot Area?
  2. Create two new variables 1) Indoor area (sum of TotalBsmtSF, GrLivArea, and 2ndFlrSF), 2) Age of the property (in Years). When looking at all of the variables (inclusive of new variables), which is the most correlated with Overall Quality?

**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 property’s Sale Price:

**R:**

model\_cleansed <- read\_csv("Model\_Cleansed.csv")

lm\_bad <- lm(SalePrice ~., data= model\_cleansed)

**Python:**

import statsmodels.api as sm

X1 = pd.read\_csv("X1.csv")

Y1 = pd.read\_csv("Y1.csv")

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)

For the rest of sub questions use the dataset generated in Q9b

* 1. Considering what you have learned about this data so far, why might we want to remove the independent variable GrLivArea from the regression?
  2. Suppose we wanted to remove any variables which are not significant (pval >0.1). How many variables would you have left?
  3. What impact on Sales Price is associated with a unit change in the Overall Quality? (for a single property with all other variables remaining constant) (answer to the nearest whole number in a real form (i.e. not in scientific form))
  4. On average, how much higher are the Sales Prices expected to be for a property with Good (“Gd”) versus a property with Poor (“Po”) Heating Quality? (answer to the nearest whole number in a real form, i.e. not in scientific form)

1. Run the follow code:

**R:**

model\_simple\_cleansed <- read\_csv("Model\_Simple\_Cleansed.csv")

model\_test <- read\_csv("Model\_Test.csv")

load("lm\_clean.rda")

**Python:**

X2 = pd.read\_csv("X2.csv")

X2\_Test = pd.read\_csv("X2\_Test.csv")

Y2 = pd.read\_csv("Y2.csv")

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

fit\_bad2 = lm\_bad2.fit()

What does the model predict for Sale Price for a property that was sold in 2010 with Lot Area of 14,267, Overall Quality of 6, Age of 64, Central Air-conditioning, Ground Living Area of 1,329, 1 Full Bathroom, Good Kitchen Quality?

1. Suppose we are interested in knowing what the sales price would be if all variables were held at their average or most common values in case of categorical values (in the Model\_Test / X2\_Test). What is the average sales price of a property across the prediction table if all other predictors are held constant? (Hint: you should set the predictors to their average value (Lot Area, Overall Quality, Ground Living Area, Full Bathrooms and Age assumed to be continuous), or their most common value if the predictor is categorical (Central Airconditioning, Kitchen Quality and Sale year)).

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

**Appendix: Data Dictionary**

***Block A***

***Table Name:*** Property.csv

***Contains:*** Details about the properties sold

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Id | Property ID |
| LotArea | Lot size in square feet |
| Street | Street: Type of road access to property   * Grvl: Gravel * Pave: Paved |
| OverallQual | Rates the overall material and finish of the house   * 10: Very Excellent * 9: Excellent * 8: Very Good * 7: Good * 6: Above Average * 5: Average * 4: Below Average * 3: Fair * 2: Poor * 1: Very Poor |
| OverallCond | Rates the overall condition of the house   * 10: Very Excellent * 9: Excellent * 8: Very Good * 7: Good * 6: Above Average * 5: Average * 4: Below Average * 3: Fair * 2: Poor * 1: Very Poor |
| YearBuilt | Original construction date |
| YearRemodAdd | Remodel date (same as construction date if no remodeling or additions) |
| RoofStyle | Type of roof   * Flat: Flat * Gable: Gable * Gambrel: Gabrel (Barn) * Hip: Hip * Mansard: Mansard * Shed: Shed |
| TotalBsmtSF | Total square feet of basement area |
| Heating | Type of heating   * Floor: Floor Furnace * GasA: Gas forced warm air furnace * GasW: Gas hot water or steam heat * Grav: Gravity furnace * OthW: Hot water or steam heat other than gas * Wall: Wall furnace |
| HeatingQC | Heating quality and condition   * Ex: Excellent * Gd: Good * TA: Average/Typical * Fa: Fair * Po: Poor |
| CentralAir | Central air conditioning   * N: No * Y: Yes |
| 1stFlrSF | First Floor square feet |
| 2ndFlrSF | Second floor square feet |
| GrLivArea | Above grade (ground) living area square feet |
| BsmtFullBath | Basement full bathrooms |
| BsmtHalfBath | Basement half bathrooms |
| FullBath | Full bathrooms above grade |
| HalfBath | Half baths above grade |
| BedroomAbvGr | Bedrooms above grade (does NOT include basement bedrooms) |
| KitchenAbvGr | Kitchens above grade |
| KitchenQual | Kitchen quality   * Ex: Excellent (5) * Gd: Good (4) * TA: Average/Typical (3) * Fa: Fair (2) * Po: Poor (1) |
| TotRmsAbvGrd | Total rooms above grade (does not include bathrooms) |
| Fireplaces | Number of fireplaces |
| FireplaceQu | Fireplace quality   * Ex: Excellent – Exceptional Masonry Fireplace * Gd: Good – Masonry Fireplace in main level * TA: Average – Prefabricated Fireplace in main living area or Masonry Fireplace in basement * Fa: Fair – Prefabricated Fireplace in basement * Po: Poor – Ben Franklin Stove * NA: No Fireplace |
| PavedDrive | Paved driveway   * Y: Paved * P: Partial Pavement * N: Dirt/Gravel |
| WoodDeckSF | Wood deck area in square feet |
| OpenPorchSF | Open porch area in square feet |
| EnclosedPorch | Enclosed porch area in square feet |
| 3SsnPorch | Three season porch area in square feet |
| ScreenPorch | Screen porch area in square feet |
| PoolArea | Pool area in square feet |
| PoolQC | Pool quality   * Ex: Excellent * Gd: Good * TA: Average/Typical * Fa: Fair * NA: No Pool |

***Table Name:*** Garage.csv

***Contains:*** Details about the garage of the properties sold

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Id | Property ID |
| GarageType | Garage location   * 2Types: More than one type of garage * Attchd: Attached to home * Basment: Basement Garage * BuiltIn: Built-In (Garage part of house – typically has room above garage) * CarPort: Car Port * Detchd: Detached from home * NA: No Garage |
| GarageYrBlt | Year garage was built |
| GarageCars | Size of garage in car capacity |
| GarageArea | Size of garage in square feet |
| GarageQual | Garage quality   * Ex: Excellent * Gd: Good * TA: Typical/Average * Fa: Fair * Po: Poor * NA: No Garage |
| GarageCond | Garage condition   * Ex: Excellent * Gd: Good * TA: Typical/Average * Fa: Fair * Po: Poor * NA: No Garage |

***Table Name:*** Transaction.csv

***Contains:*** Details about the sale of the properties

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Id | Property ID |
| MoSold | Month Sold (MM) |
| YrSold | Year Sold (YYYY) |
| SalePrice | The property's sale price in dollars |