**Scenario:**

The case for project 1. You are still working as a consultant for Ms. Roices and this time she is interested in knowing the factors that determine the price of real estate in LA area. Your work in Project 1 has given her some idea about the distribution of the real estate prices in different areas of LA. However, she now wants a predictive model that can allow her to predict house prices based on the features of the house so that she can use that information in her real estate investment strategy. The specific tasks to complete are:

1. Create dummy variables for the house types. Use 1 for single family residence and 0 for Townhouse. *(2 points)*
2. Create dummy variables that would allow to investigate the effect of the location of the properties in LA. You can use the information on the website shown below to get the regions for your properties or you can make up the regions on your own. Make sure to aggregate the data in a way so that they make sense for price predictions (e.g. you should not clump beach cities and inner valley properties together) and keep the total number of regions to about 5.*(8 points)*



<http://cretscmhd.psych.ucla.edu/healthfair/SPAs%20-%20page%20and%20PDFs/SPAsmain.htm>

1. Use descriptive analytics to summarize the data. The should include a pivot table to show the 5-number summary for house price and price/sqft for each of the regions as well as correlation analysis and pairplots (through the Seaborn package in Python or Tableau) between all the variables. *(5 + 5 + 5 = 15 points)*
2. Develop a simple linear regression models that can be used to predict the house price for a given house size (Sq ft). *(5 points)*
3. Develop another simple linear regression models that can be used to predict the price/sqft for a given house size (Sq ft). *(5 points)*
4. Discuss your findings from d and e. What kind of insights do you gain from the models? *(5 points)*

In part D when we examine the R Square value it is roughly 38% which is below 50% which can denote that the model is weak. Furthermore, the P-value in this scenario for Square Feet is far below 5% which means we can confidently accept the alternative hypothesis. This leads us to the equation of Price = 722948.7367 + 564.4787\*SQUARE FEET. The equation is interpreted as follows; for every unit increase in SQUARE FEET the price increases by roughly $564.48.

In Part E when we examine the R Square value it is far below 50% meaning that the model is incredibly weak. Next the P-Value for the SQUARE FEET as it pertains to price per square foot is roughly 28% which is well above 5% meaning that we fail to reject the null hypothesis, thus SQUARE FEET does not have a real difference on Price per SQFT. When we look at the equation it is notable that SQUARE FEET is not much of a factor here.

1. Based on the pairplots created in part c, identify other possible variables that can impact the house price and price/sqft along with the nature of the relationships. *(5 points)*

Price can be impacted by BATHS, BEDS, and SQUARE FEET positively whilst it can be impacted by Distance from ocean negatively.

Price Per Sqft can be impacted by Distance from ocean and AGE both negatively.

1. Develop a multiple linear regression model that can be used for predicting house price for using all possible independent variables. Discuss the quality of the models, significance of the various variables and the applicability of the model. *(8 points)*

When investigating the variables I first explored their relationship to Price. Doing so I used the Linest model excel file. BEDS had the least amount of skew when it was Sq-Root, BATHS had the least amount of skew when it was Linear, SQUARE FEET had the least amount of skew when it was Linear as well, and Distance from ocean had the least amount of skew when it was Sq-Root.

I decided to transform the variables in the Linest model to get a more accurate model. The BEDS variable had the highest p-value over 5% and so eliminated it first followed by BATHS which had the next highest. This left me with SQUARE FEET and Distance from ocean. The adjusted R-Square was close to 59% which was better than if I had not transformed the variables prior. This adjusted R-square is indicative of this model being fairly effective in attaining the home price value.

The model I developed for the Price utilizing all the possible aforementioned variables (i.e. BATHS, BEDS, SQUARE FEET, and Distance from ocean) was Price = 1864933.85\*LN(SQUARE FEET) – 224182.51\*Sq-Root(Distance from ocean) – 11559356.42

1. Develop a multiple linear regression model that can be used for predicting price/sqft for using all possible independent variables. Discuss the quality of the models, significance of the various variables and the applicability of the model. Discuss any possible multicollinearity issue that can exist in the model. *(7 points)*

When developing the model, I examined the relationship between the possible independent variables (i.e., AGE and Distance from ocean) and price per sqft. I noticed that transforming the Distance from ocean variable into Sq-root resulted in less skew in the data. Following this I utilized the linest model to craft the multiple regression model. The variance inflation factor for both variables is a very low level of 1 which indicates there would be no multicollinearity concerns.

Overall, the adjusted R-square of the model is very low at roughly 10% however when applying the Sq-Root transformation to Distance from ocean improved the R-square to 12%. This tells me that the model is weak and not reliable for attaining the price per sqft value. Although the variables themselves have a p-value that is well below 5% indicating that they have an impact on the price per sqft even though the model itself is not a strong one. Additionally, the CP was 1 and I ended up with 2 variables.

The equation was:

Price per sqft = 2.68\*AGE – 132.59\*Sq-Root(Distance from ocean) + 1077.68

1. Using either the LINEST model or the regression models available in Python or R, develop the “best” possible multiple regression models for predicting the house price as well as for predicting the price/sqft (two separate models). Remember that the best model is the one with minimal number of variables that are significant without sacrificing the adjusted R-square value. Briefly explain your rationale for choice of the variables and the quality of the model. *(20 points)*

For “Price” I gathered all of the independent variables that could possibly have an effect on price so as to leave no stone unturned from the data I gathered. My method for choosing the variables was partly by the correlation table and mostly by my own intuition. The correlation table I made for Price listed BEDS BATHS and SQUARE FEET as all strong candidates for affecting the price. However, I figured that when taking all of the regions dummies into account such as Metro San Fernando South Bay West etc. that I would find more variables affecting the price. As I was certain that Distance from ocean, and LOT SIZE must be a factor and perhaps AGE as well.

Following this I checked the relationship of these independent variables and how skewed they might be. Observing the independent variables, I applied the transformations that resulted in the least amount of skew. Now, I moved on to playing around with the variables and checking how it affected Price. The South region had no effect whatsoever, so I immediately removed it. The South Bay had a massive p-value as well as an incredibly large VIF which signified that there was multicollinearity at play here leading me to remove the South Bay and then the San Fernando region as well for similar issues. Afterwards I removed AGE as it had very minimal affect on the Adjusted R-Sq value. Now, I was left with the following variables: BEDS, BATHS, SQUARE FEET, LOT SIZE, Distance from ocean, House Type, Metro, and West. My adjusted R-Sq value was roughly 68% indicating a reliable model for calculating Price, the p-value for all of my variables left were well below 5% and the VIF for each of my variables in the final model was well under 5, thus multicollinearity should be of no concern here.

My equation for calculating the price was:

Price = -645739.2\*Sq-Root(BEDS)+404144\*BATHS+1118872\*LN(SQUARE FEET)\*127839.64\*LN(LOT SIZE)-143005.43\*Sq-Root(Distance to ocean)+842836\*House Type Dummy+1015432.03\*Metro+611894.99\*West-8127039.568

Now on to price per sqft, my method here was more calculated as I had an idea what should affect price per sqft as well as the information I could surmise from my correlation table. The independent variables for LOT SIZE, Distance from ocean, AGE, House Type, and all the regions seemed like a reasonable and logical collection of variables to create a model for calculating price per sqft. Next I checked the variable relationships and examined the skewness of the data checking whether they are positively or negatively skewed. Following this I began playing with the variables in the linest model. I transformed the Distance from ocean due to its skewness and its overall affect on the Adjusted R-sq. Then, I removed variable for the West as it had an incredibly high VIF and p-value. I removed LOT SIZE for it too had a high p-value and thus this left me with Distance from ocean, AGE, House Type, Metro, San Fernando, and South bay. My Cp was 5 and I was left with 6 variables. Each variable had a VIF well below 5 and a p-value all well below 5%. My adjusted R-Sq which was weak was roughly 28% which is up from my previous model that had around 12%.

My equation was:

Price per sqft = 9.84\*Sq-Root+1.83\*Age+180.68\*House Type Dummy + 364.86\*Metro-257.00\*San Fernando-303.59\*South Bay + 1072.23

1. Does the distance from ocean play significant role in determining the house price or price/sqft? Does it vary from one region to the next? Is there a cut-off point beyond which the distance does not really matter? Briefly explain your answers. *(5 points)*

The distance from ocean does play a role in determining the house price and the price/sqft as indicative of the scatter plots, I made in excel. However, this role is dependent on the region as not all regions share this same distinction. When it is concerning the price the South Bay region is affected by the distance from ocean and there is a noticeable drop in price the further you move away from the ocean until you are roughly 20 miles out. Now, in the case of price per sqft the West and South Bay regions are definitely affected by the distance from the ocean. The West and South Bay regions both have a cut off of around 9 miles.

1. In the areas where the distance from ocean is a factor in determining the house price, what is the rate of change in the house price or price/sqft for change in the distance from the ocean? Explain your answer with the relevant results from your regression analyses. *(5 points)*

I was able to calculate the rate of change by performing a simple regression on the regions price and price per sqft data as it is affected by distance from the ocean.

For the South Bay region for every increase in unit of distance from the ocean the Price will decrease by $56,286.68 and the Price Per Sqft will decrease by $7.19.

For the West Region the Price Per Sqft will decrease by $16.05

1. What conclusions and recommendations can you derive from your analysis? What is your final recommendation to your client? What other independent variables could be included in the model? *(10 points)*

My recommendation to Ms. Roices would be to consider the following factors that pertain to price in LA real estate market: beds, baths, square feet, lot size, distance from ocean, home type, and region. Now, when concerning the price per sqft it should be noted that the beach cities in the South Bay and West regions are affected by the distance from the ocean within 10 miles afterwards the distance from ocean is not of a concern. In an effort to enhance this analysis the following independent variables could have been included such as median or average household income, crime rate by city and or region, number of homeless encampments, and number of schools.