Name: Jatin Padyar

Assignment-based Subjective Questions

1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

⇒ **Following categorical variables are present in the dataset: season, month, holiday, weekday, workingday, year, weathersit.**

From EDA we can infer the following:

- **We can infer that popularity grew over the years from 2018 to 2019 which is indicated by increase in rental count.**
- **Average number of rentals were higher when weather was clear followed by misty and then light rain/snow climate.**
- **Average number of rentals is highest in fall season and least in spring season.**
- **No particular trend observed for demand w.r.t month but the median value of demand was highest in the month of July and least in January.**
- **No particular trend observed for demand w.r.t week day but the avg value of rentals was least on Sunday.**
- **No particular trend observed for demand w.r.t workingday as avg value of rentals is approx. same for both working and nonworking day.**
- **avg value of rentals is higher for holiday compared to its counterpart.**

But during regression modelling some of the categorical variables may become insignificant. Thus, categorical variables, more specifically dummy variables present in final model are as follow: year, spring, winter, mist, light rain/snow, July, September and Sunday.

- **For dummy variables having positive slopes, if true i.e., its value equals to 1 then target variable increases by magnitude of its slope, provided all other variables are kept constant. e.g., If September demand will increase. Such categorical variables are: year, September, winter.**
- **For dummy variables having negative slopes, if true i.e., its value equals to 1 then target variable decreases by magnitude of its slope, provided all other variables are kept constant. E.g., If light rain/snow, demand will decrease. Such categorical variables are: Sunday, July, mist, spring, Light rain/snow.**

---

2) Why is it important to use drop_first=True during dummy variable creation?

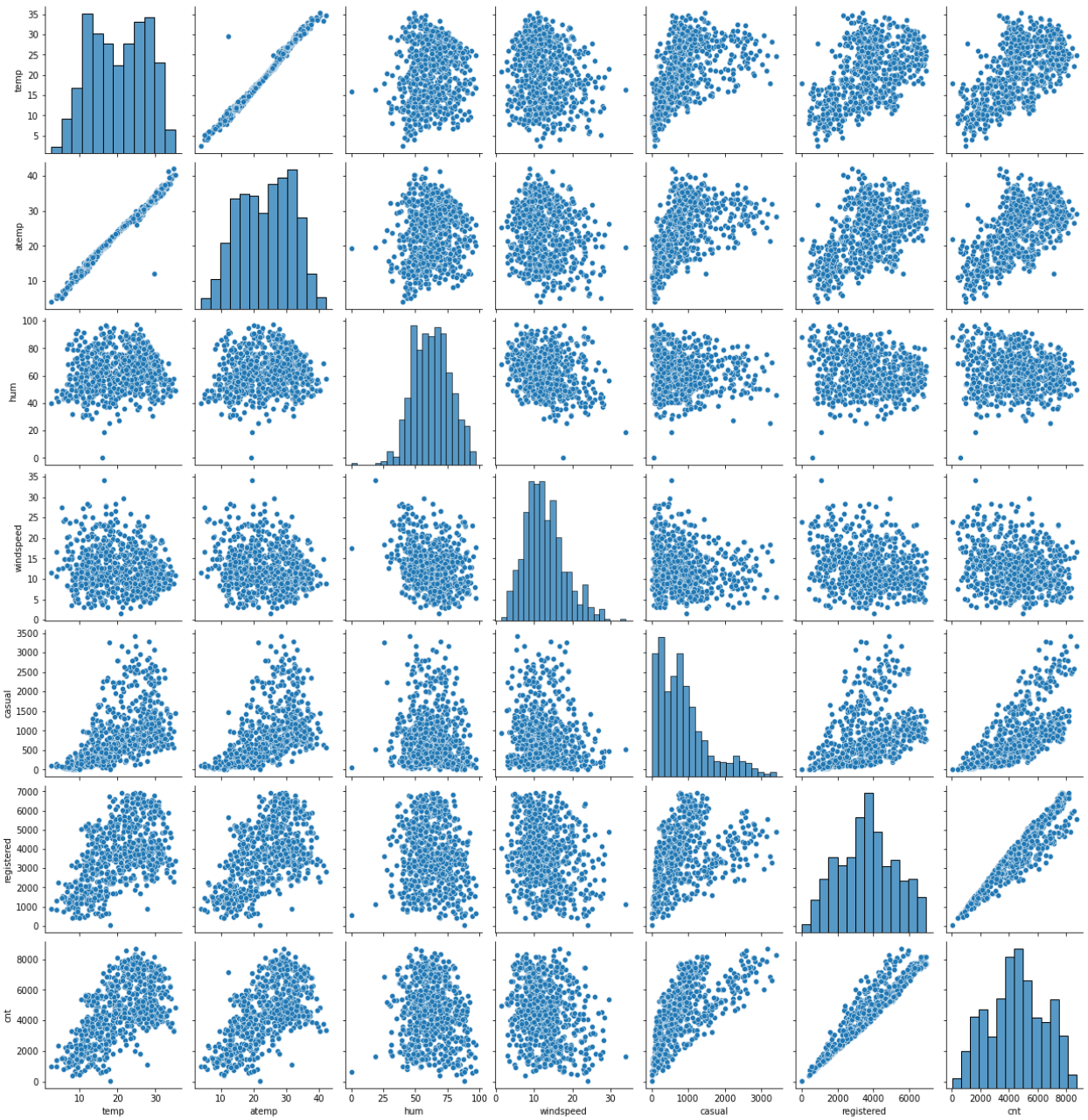
⇒ **Suppose categorical variable has k levels. These k levels can be represented perfectly with the help of (k-1) dummy variables instead of k dummy variables. When we use drop_first=True during dummy variable creation, it avoids creation of extra dummy variable which is redundant thus increasing efficiency of the model. Another main reason is to avoid dummy variable trap. If we don't drop one of the levels, it causes multicollinearity with other levels which may reduce the performance of linear regression model. Therefore, Since, one-hot-encoding introduces multicollinearity, we drop one of the levels of categorical variable.**

---

3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

⇒ **Refer the pair-plot below corresponding to target variable 'cnt':**

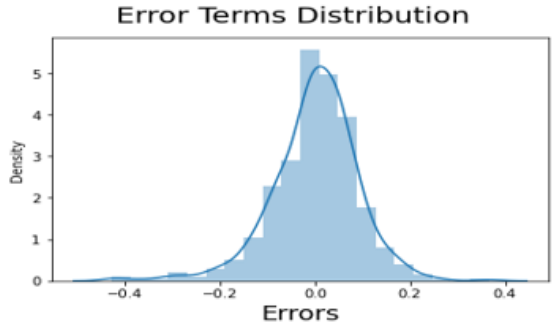
- **If we consider all the variables then, registered has the highest correlation with target variable, followed by casual, temp and atemp respectively.**
- **correlation with target variable is highest for variables temp and atemp, after dropping registered and casual features to avoid data leakage.**

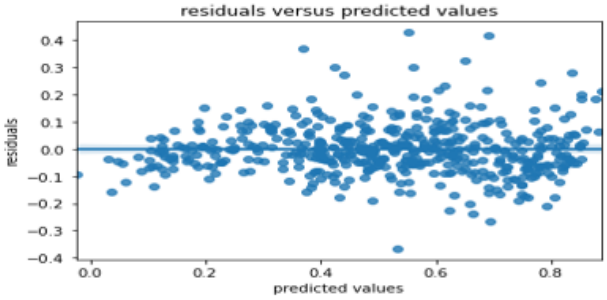


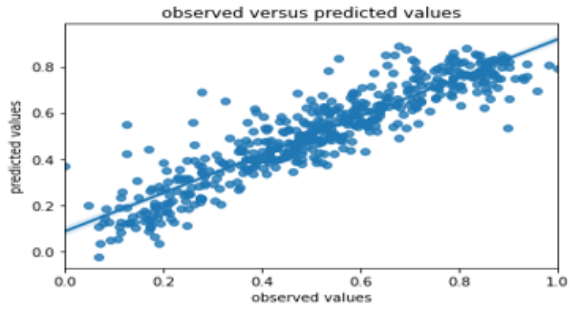
4) How did you validate the assumptions of Linear Regression after building the model on the training set?

⇒ **With the help of various plots e.g., scatterplot, distplot, Q-Q plot, regplot, we validated initial assumptions using visualization.**

- **assumption of Linear relationship between dependent and independent variables is validated using pair plots (more specifically scatter plots of target variable against each predictor).**
- **With the help of distplot i.e., distribution of error terms we validated assumption of residuals being normally distributed with mean centered at zero. Same can be validated with Q-Q plot.**
- **With the help of regplot / scatter plot i.e., by plotting predicted values of training data against residuals, we checked if there are any patterns present. This Randomness validated assumption of error terms being independent of each other.**
- **With the help of regplot / scatter plot i.e., by plotting predicted vs observed values for training data, we checked whether or not error terms have constant variance.**
- **With the help of correlation heatmaps and VIF values we checked for multicollinearity between predictor variables.**







---

5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

⇒ **Following are the top 3 features contributing significantly towards demand of shared bikes:**

- **Temperature with beta coefficient of 0.45**
- **weathersit_Light rain/snow with beta coefficient of -0.286574**
- **year with beta coefficient of 0.235098**

General Subjective Questions

1) Explain the linear regression algorithm in detail.

- **Linear regression falls under supervised learning method where past data is used for building the model. In Regression, the output variable to be predicted is a continuous variable. Linear regression is one of the types of regression and it is used when there is a linear relationship between dependent and independent variables.**
- **In simple linear regression (SLR), there is only one independent variable present. In SLR we try to fit a best straight line through the data.**
- **In multiple linear regression (MLR), there are more than one independent variable present. In MLR we try to fit a best hyperplane through the data.**
- **General equation for linear regression is as follow:**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i$$

Where, Y = dependent variable, β_0 = y-intercept, β_i = slope for X_i and X = independent variable.

- **We find the best fitted line/hyperplane by finding these beta coefficients for which cost function i.e. residual sum of squares is minimum. For finding optimum solution for cost function, we use gradient descent approach.**
- **Following are the assumptions of linear regression:**
  - **There is a linear relation between dependent and independent variable.**
  - **Error terms are normally distributed with mean zero**
  - **Error terms are independent of each other.**
  - **Error terms have constant variance (Homoscedasticity).**
  - **Predictor variables should not have Multicollinearity in case of MLR.**
- **Linear regression guarantees interpolation of data and not the extrapolation.**
- **Linear regression is a parametric model i.e., in simple terms it can be described using a finite number of parameters.**

---

2) Explain the Anscombe's quartet in detail.

⇒ **The Anscombe's quartet was constructed by the statistician Francis Anscombe to demonstrate the importance of visualizing the data and the effect of outliers on statistical properties.**

⇒ **Anscombe's quartet contains four datasets that have identical simple statistical summary e.g., mean, variance etc. yet have very different distributions and appear very different when visualized using plots.**

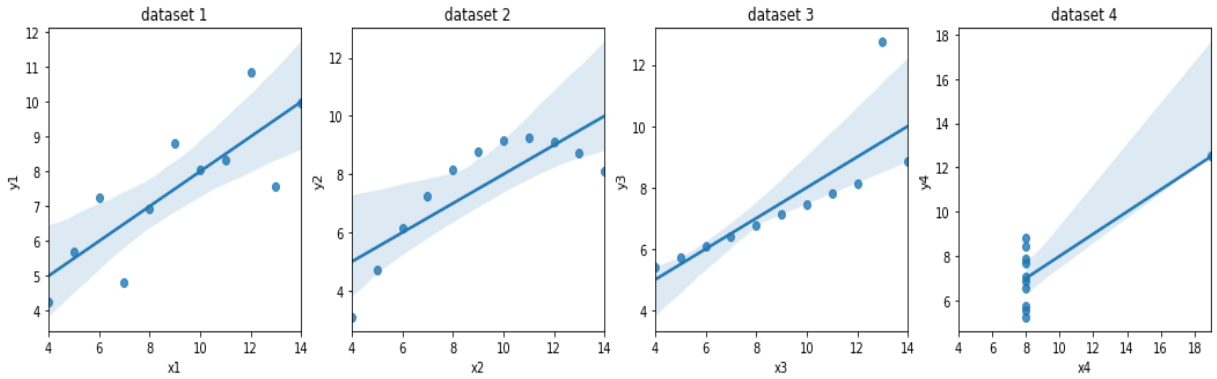
Following are the four datasets that Anscombe came up with: (source: geeksforgeeks.org)

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

⇒ **If we find simple stats for above four datasets, it will be identical e.g.,**

- **Mean and standard deviation of X comes out to be 9 and 3.32 respectively for each dataset.**
- **Mean and standard deviation of Y comes out to be 7.5 and 2.03 respectively for each dataset.**
- **Correlation between X and Y comes out to be approx. 0.816 for every dataset.**

⇒ **From above quantitative simple stats summary, one would assume that all four datasets are same. But that would be a mistake.**



⇒ **When we visualize each dataset, we can clearly see that each dataset has different distribution.**

- **1st plot shows linear relation between x and y.**
- **2nd plot shows nonlinear relation between x and y.**
- **Even though 3rd plot shows linear relation between x and y, its coefficients and thus regression line would be different from that of 1st plot due to presence of an outlier.**
- **4th plot represents the case when there is one high-leverage point and it is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.**

⇒ **Basically, Anscombe wanted to counter the early statistical belief that indicated numerical calculations are exact, but graphs are rough, which is clearly not the case.**

---

3) What is Pearson's R?

⇒ **Pearson's R is also called as bivariate correlation coefficient is a measure of linear correlation between two variables.**

⇒ **It is calculated by dividing covariance of two variables by product of their standard deviation.**

- **Pearson coefficient (r) = COV(X,Y)/(std(X)*std(Y))**

⇒ **Its value lies in the range of [-1,1]**

⇒ **Magnitude represents strength of correlation. Higher the magnitude stronger the correlation.**

⇒ **Whereas sign represents type of correlation (positive or negative). Positive correlation indicates that as one variable increases, other variable also increases. negative correlation indicates that as one variable increases other variable decreases.**

- **r = 1 indicates perfect positive correlation i.e., all the data points lie on a straight line having positive slope.**
- **r = 0 indicates no correlation between two variables.**
- **r = -1 indicates perfect negative correlation i.e., all the data points lie on a straight line having negative slope.**

⇒ **It should be noted that Correlation does not imply causation.**

---

4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

⇒ **Feature Scaling is a method to standardize the independent features present in the data in a same range/scale so that no variable is dominated by the other.**

⇒ **Most of the times dataset has variables with hugely varying scales/ranges/units. Most of the times, underlying ML algorithm uses Euclidean distance which considers magnitude of the variables. If data points are not scaled properly i.e., variables have large variation in range then the features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes.**

⇒ **Scaling also improves efficiency of gradient descent method which is used in core of many ML algorithms for optimizing cost function.**

⇒ **So, in simple terms, even though scaling does not affect goodness of fit of the model, we should do scaling for following reasons:**

- **To achieve faster convergence of gradient descent**
- **To make useful interpretations from the model as without scaling some of the beta parameters may get too big or small as compared to others and it then becomes harder to make sense.**

⇒ **Standardized scaling transforms the data such that the resulting distribution has mean of 0 and a standard deviation of 1. Normalized scaling affects the value of dummy variables. If we are going to use ML algorithm or statistic technique which assumes data is normally distributed (e.g., t test, ANOVAs etc.) then we should use standardized scaling.**

- **Standardized scaling uses mean and standard deviation for scaling.**
- **Standard scaling:** $x = \frac{x - mean(x)}{sd(x)}$

⇒ **Normalized/MinMax scaling transform the data such that the feature values are within a specific range e.g. [0, 1]. Even after MinMax scaling values of dummy variables remains same. Where distance between the data points is important (e.g., SVM, KNN), we generally use MinMax scaling.**

- **Min and Max values of feature are used for scaling.**
- **Normalized scaling (x) =** $\frac{x - \min(x)}{\max(x) - \min(x)}$

---

5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

⇒ **VIF calculates how well one independent variable is explained by all other independent variables combined i.e., how well a predictor variable is correlated with all other variables, excluding target variable.**

⇒ **It is calculated as follow:** $VIF_i = \frac{1}{1 - R_i^2}$ **; where i refers to ith variable which is being represented as a linear combination of rest of the independent variables.**

⇒ **From above equation VIF will become infinite when denominator becomes zero i.e., when** R_i^2 **is equal to 1.**

⇒ R^2 **equals to 1 is the indication of perfect multicollinearity.**

⇒ **Thus, in case of perfect multicollinearity, where given independent variable can be perfectly explained/represented as linear combination of other independent variables, value of VIF becomes infinity.**

---

6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

⇒ **The Q-Q plot (quantile-quantile plot), is a plot that helps us to check if a set of data came from some theoretical distribution such as a Normal or exponential distribution.**

⇒ **It is a scatterplot created by plotting two sets of quantiles against one another.**

⇒ **If all the data points lie on a straight line, we can assume that both sets of data came from normal distribution.**

⇒ **This property of Q-Q plot makes it important in linear regression while doing residual analysis.**

⇒ **One of the initial assumptions of linear regression is that error terms are normally distributed. We can validate this using Q-Q plot by plotting theoretical and sample quantiles for residuals in Q-Q plot and checking whether datapoints lie on the straight line or not.**