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?**

Year, working day, holiday, weather situation , month and weekday are categorical variables in given dataset
Dependent variable is total rental of bikes, please refer Figure 1 for plots of all Categorical variables with bike rental

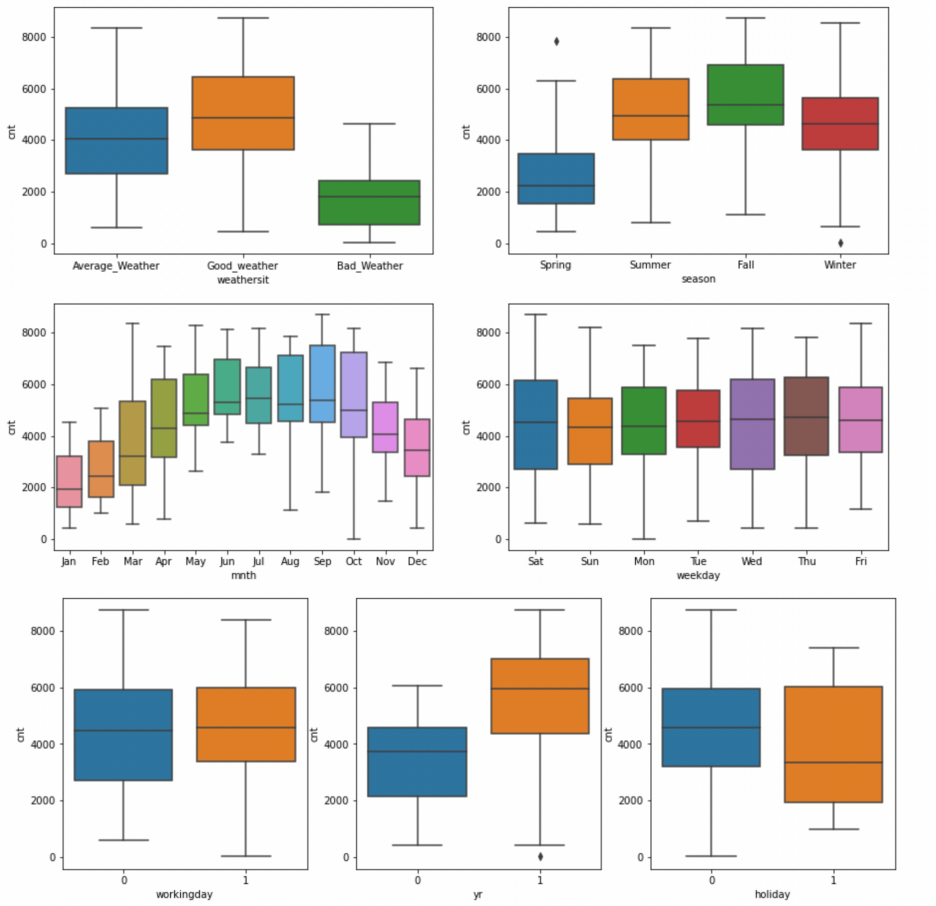


Figure 1: Categorical variable Vs Dependent variable(Bike rental)

Inferences of categorical variable on bike rentals :

1. Bike rentals is largely dependent on weather situation if it goes worse rentals are lower and vice-versa, also during extreme weather conditions no one rents bikes
2. Bike rentals is also seasonal dependent increased for summer and further increased during fall but decreases during winter
3. Bike rentals increase from Jan to Jul and from Jul to Oct is the best season for bike rentals then rentals dip in Nov, Dec
4. Bike rentals seems not much dependent on day of week
5. Bike rentals demand increased with time (median of rentals in year 2019 is higher compared to year 2018)
6. In holidays demand for rental is slightly lower
7. Working or non-working day does not affect rentals much

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

"Drop_first = True" will help in reducing first column created during dummy variable creation, thus reducing correlations amongst the created dummy variable. We can drop first column as the dropped column can be expressed as all other dummy variables = "0" i.e similar to the dropped first column value = 1.

The key idea behind dummy encoding is that for a variable with, say, 'N' levels, you create 'N-1' new indicator variables for each of these levels, refer below example.

Suppose we have categorical variable " relationship status" and have 3 categories, so we created 3 dummy variables , dummy variable is kept as binary "1" or "0". Refer Figure 2 where " single" is expressed as 1(single),0(in a relationship),0(married)

Relationship Status	Single	In a relationship	Married
Single	1	0	0
In a relationship	0	1	0
Married	0	0	1

Figure 2: dummy variable with (3 level)

If we dropped first variable "single" is expressed as other dummy variables as "0" i.e 0(in a relationship),0(married) refer Figure 3

Relationship Status	In a relationship	Married
Single	0	0
In a relationship	1	0
Married	0	1

Figure 3: dummy variable with (2 level) with drop_first = True

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

Temperature has highest correlation with target variable.

Correlation coefficient is 0.63 and highest amongst numerical variables, it is positive relationship thus if temperature increased bike rentals also increased.

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

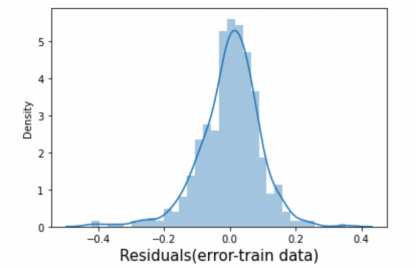
Assumptions of linear regression validated:

I. There is linear relationship with variables

We have seen from pair plots that there is linear relationship between independent variables and target variable(rentals)

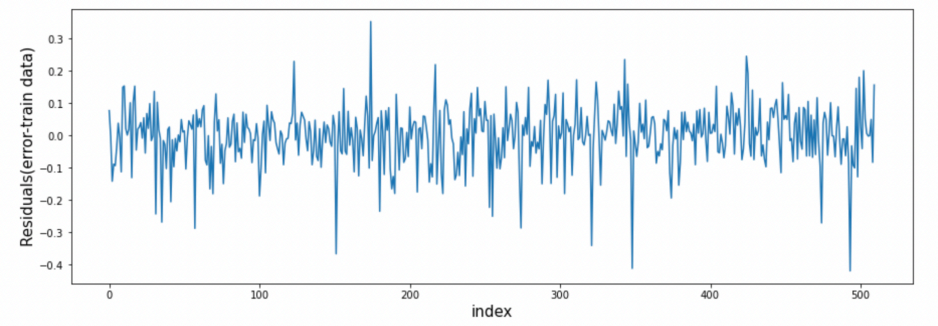
II. Error terms are normally distributed

Below plot of distribution of error(residuals) is having normal distribution



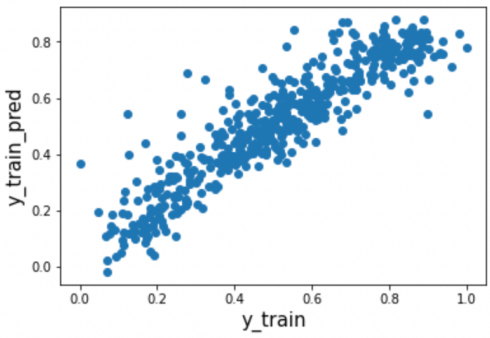
III. Error terms are independent of each other

Below plot shows error terms are independent of each other, do not follow any pattern



IV. Error terms have constant variance

Below plot on train data shows error terms have constant variance (homoscedasticity)



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

Top 3 feature explaining demand of shared bikes rentals are

1) Temperature
2) Year
3) Weather situation-3 (light snow/rain)

Above 3 have highest coefficient for linear equation

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

linear regression is a method of finding the best straight line fitting to the given data, i.e. finding the best linear relationship between the dependent(y) and independent variables/features (x). It is based on y = mx + c equation.

- y is dependent variable
- x is independent variable
- m is slope of line , i.e change in value of y for 1 unit change in x
- c is intercept , i.e value of y when x = 0

linear regression algorithm is a supervised machine learning algorithm that's finds best linear-fit between dependent(y) and independent variables(x).

- Best line is fitted using least square method where residual sum of squares is minimized.
- In order to validate if coefficient are significant or not we do hypothesis testing.
- R2 measures how much of the variance in data is captured by the model.it varies between 0 and 1. Higher the value i.e close to 1 of R2 better the model
- Finally F- statistic is used to understand overall model fit. It the probability of F statistic is less then we can say that the overall model fit is significant.

Assumptions of linear regression are:

I. There is linear relationship between dependent and independent variables
II. Error / residual are normally distributed
III. Error / residuals are independent of each other
IV. Error / residuals have constant variance

Regression is mainly divided in to two types simple and multiple linear regression

Simple linear regression:

In Simple Linear Regression the no of independent/predictor variables is 1. Model fits a straight line between X & y.

$$y = \beta_0 + \beta_1 X_1$$

Multiple linear regression:

In Multiple Linear Regression the no of predictor/independent variables is more than 1.
Model fits a hyper plane between various values of X & y.

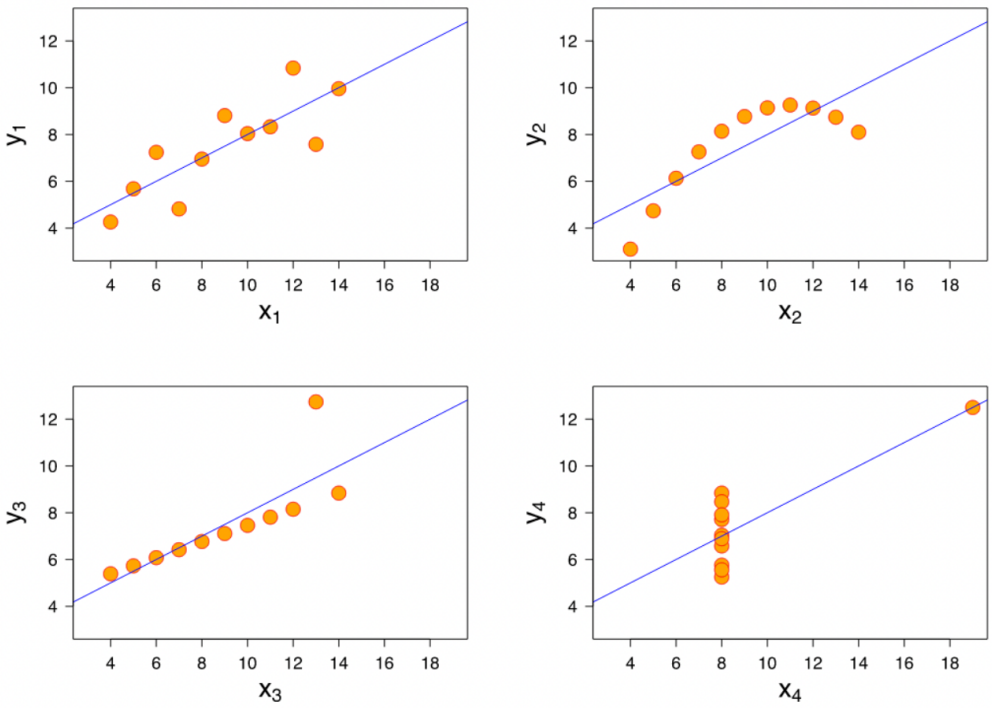
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

For multiple linear regression, feature selection is important step. All the features need not be taken in input for model we need to address various factors/issues like:
Multi-collinearity: dependence of one variable on rest other variable needs to be considered. VIF is used to understand multi-collinearity.
Dummy variables: creating dummy variables for categorical variables is important in order to include that variable in the regression model.
Feature Scaling: it is important in linear regression as it increases ease of interpretation and results in faster convergence
For multiple linear regression model adjusted R2 is used to compare models.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is group of four datasets(each having 11 (x,y) pairs) constructed by statistician Francis Anscombe in 1973 to demonstrate importance of visualizing data, effect of outliers and other influential observations on statistical properties.

He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough".



Dataset I: set of points which follow linear relation with some variance.
Dataset II: does not follow linear relation ship instead would have fit a curve.
Dataset III: looks like a tight linear relationship between *x* and *y*, except for one large outlier.
Dataset IV: x is constant for all y except one data point which is outlier

Surprising all 4 data sets have same summary statistics shown below.

Property	Value
Mean of x	9
Sample variance of x : s_x^2	11
Mean of y	7.50
Sample variance of y : s_y^2	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression : R^2	0.67

Computing summary statistics or staring at the data wouldn't have told us any of these stories. Instead it's important to visualize the data to get a clear picture.

3. What is Pearson's R?

Pearson's R or correlation coefficient is measure of degree of linear relationship between two variables X & Y, it is given by formula:

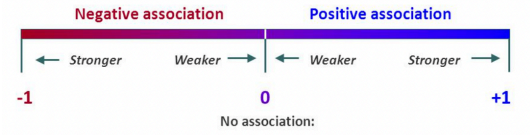
$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

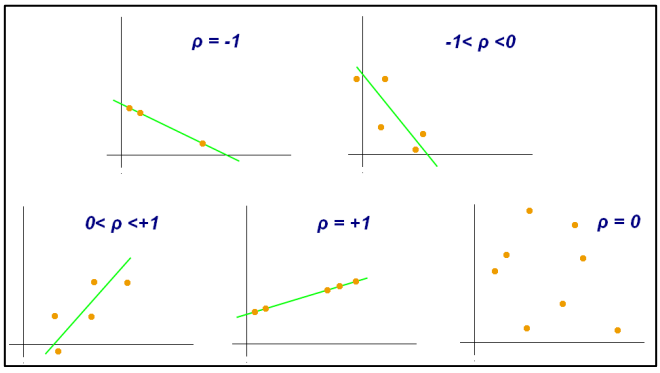
where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

- Value of R ranges from -1 to +1
- The sign of R indicates direction of :
  - Positive Value: increasing one variable results in increase of other variable
  - Negative Value: increasing one variable results in decrease of other variable and vice-versa
- The size of R indicates strength of relationship between two variables:
  - Value of R close to +1 or -1 shows strong relationship
  - Value of R close to 0 indicates weak relationship

Below examples give and idea about R value:



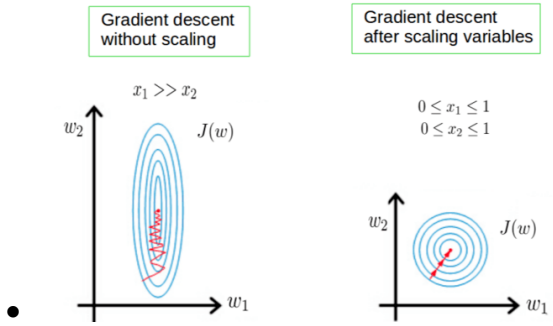


4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units

Scaling is performed as:

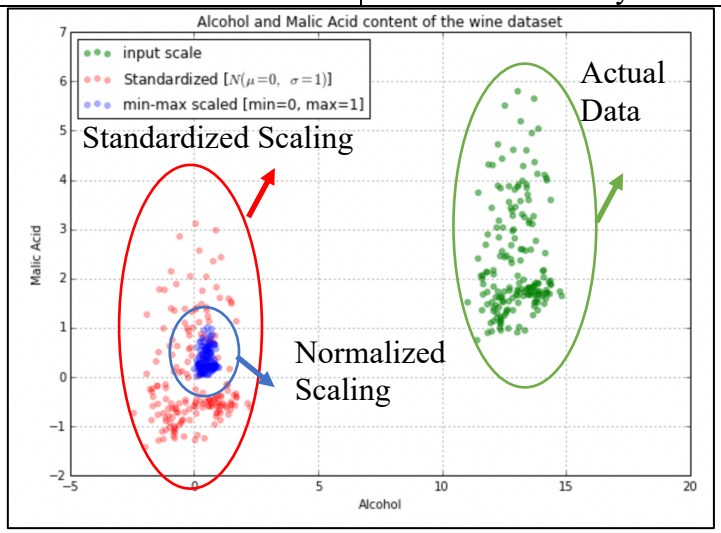
- it increases ease of interpretation: if absolute value of coeff of 2nd variable is more than 1st variable then we can say predicted variable is more dependent of 2nd variable etc
  for scaled data if weight= 0.4 * height + 0.2 * bone density
  we can say weight is more dependent on height rather than bone density
- results in faster convergence for example convergence of gradient descent methods.



- If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. Ex: If an algorithm is not using the feature scaling method then it can consider the value 100 gm to be greater than 5 kg but that's actually not true

Difference between standardization and normalization:

Normalization scaling	Standardization scaling
Minimum and maximum values are used for scaling	Mean and standard deviation are used for scaling
Xi = (Xi- Xmin)/(Xmax – Xmin)	Xi = (Xi – Xmean)/(Std. dev.(X))
Scaled values are in range of 0 to 1	Scaled values have mean = 0 and standard deviation = 1
It is effected by outliers	It is less effected by outliers



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

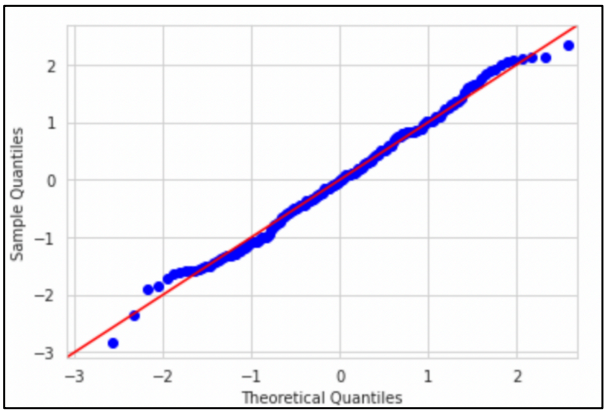
The VIF of a variable indicates the strength of the linear relationship between the variable and the remaining independent variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. It is measure of multicollinearity, it is given by:

$$VIF = 1/(1-R2)$$

Value of VIF for a variable can become infinite if there is a perfect correlation between that variable and other variables, which means that the variable can be exactly represented as a linear combination of other variables with a R2 value of 1.

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

Q-Q plot(quantile – quantile plot) is scatter plot of quantile of first dataset against quantile of second dataset. By quantile we mean fraction of points below a value ex: 0.4 quantile value tell you that 40% of data are below that particular value.
Q-Q plot is a graphical tool for comparing two distributions by plotting their quantiles against each other . if the both datasets have same distribution they fall on 45 Deg line. Below is an example of Q-Q plot which have same distributions:



It is used to check following scenarios:
If two data sets

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behavior

Importance of Q-Q plot in linear regression:

- To check if residual are normally distributed or not
- If training and test data set provided separately , we can infer if both datasets are from population with same distribution or not