Assignment-based Subjective Questions

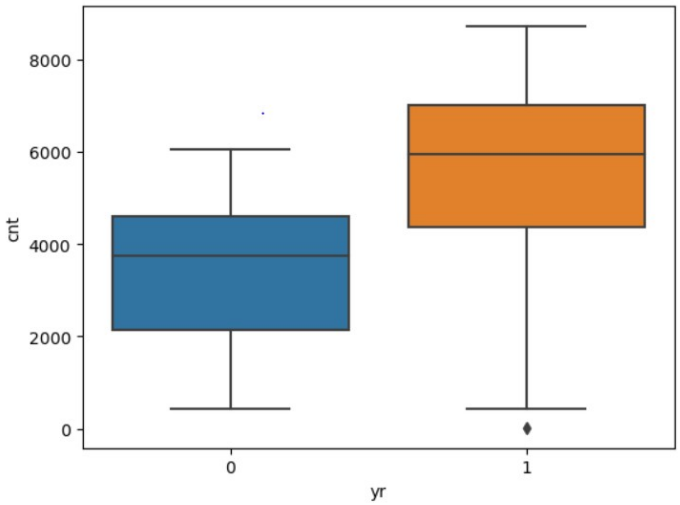
Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans:
From the dataset provided 'day.csv' there are 7 categorical variables which are finally selected for initial model development out of which 2 are converted to dummy variables as follows:

1. yr
2. mnth
3. holiday
4. weekday
5. workingday
6. season: (Converted to dummy variables)
   - spring
   - summer
   - winter
   - fall(dropped)
7. weathersit: (Converted to dummy variables)
   - weather-t1(dropped)
   - weather-t2
   - weather-t3

Based on analysis from categorical variables we can infer following observations:

(1) yr:
As per observation from distribution with respect to target variable, 'yr' (year 2019=1, year 2018=0)category shows as follows:



- The more number of BikeShaing bookings are happened in year 2019(1) compared with year 2018(0).
- The trend shows that BikeSharing bookings are increasing every year due to popularity and usage.

(2) mnth:

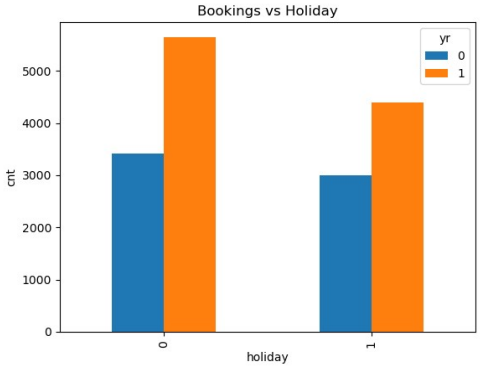
As per observation from distribution with respect to target variable, ‘mnth’ category shows as follows:

From the above figure we can infer following points with respect to dependent variable:

- More number of BikeSharing bookings happened in September-2019 month, June-2018 month. This might be due to **lock down** in 2019.
- And there is an increasing pattern of bookings from January to mid of year and decrease in bookings from mid to end of year. This might be due to people going for **vacation** mostly in months from October to March.

(3) holiday:

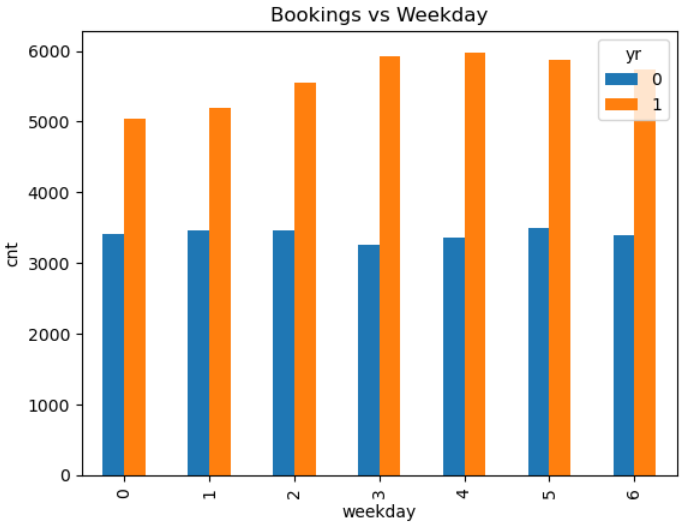
As per observation from distribution with respect to target variable, ‘holiday’ category shows as follows:



- From the above plot we can infer that many people are using shared bikes on non-holiday, might be using bike sharing for office commute.

(4) weekday:

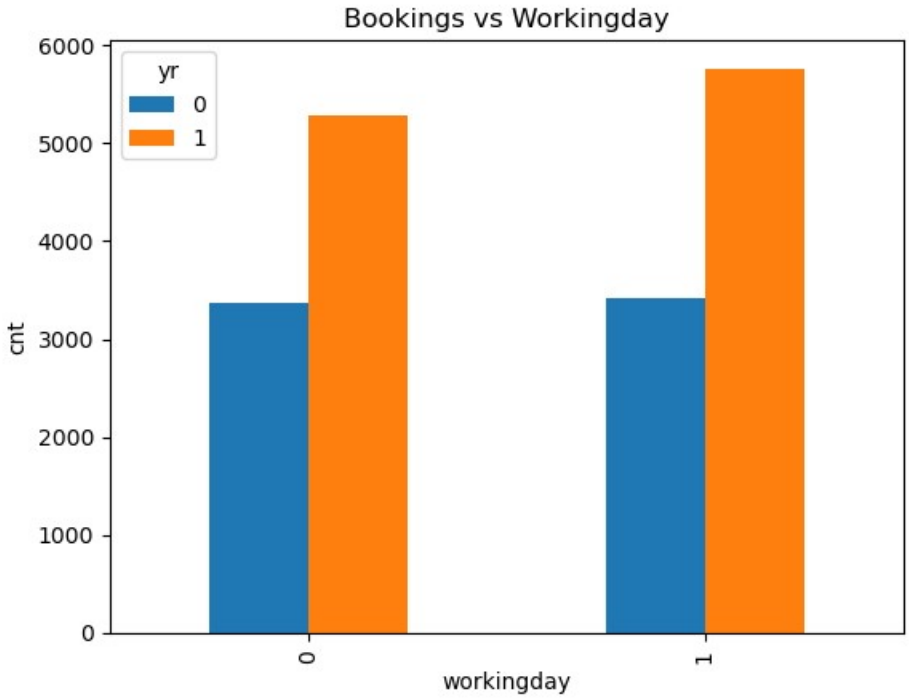
As per observation from distribution with respect to target variable, ‘weekday’ category shows as follows:



From the above plot we can observe that most bike sharing bookings are happened in mid of the week in 2019 and marginally less bookings in mid-week in 2018. this is because of hybrid work model in 2019.

(5)workingday:

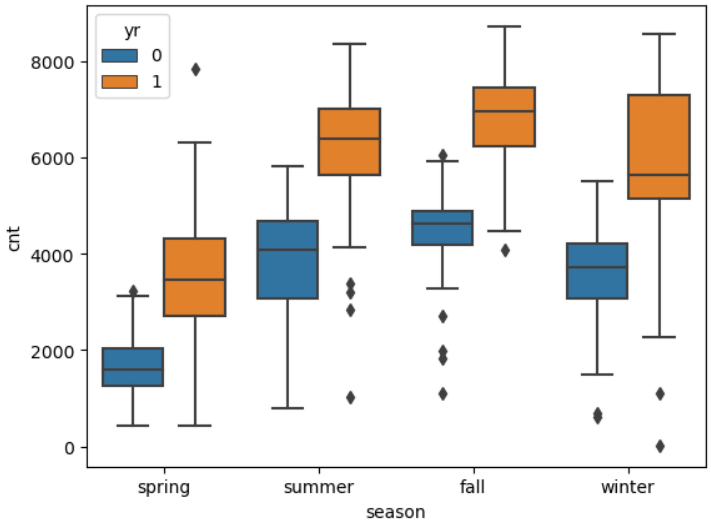
As per observation from distribution with respect to target variable, ‘workingday’ category shows as follows:



From the above plot we can observe that there is small increase in bookings on workingdays compared with non-working days.

(6) season:

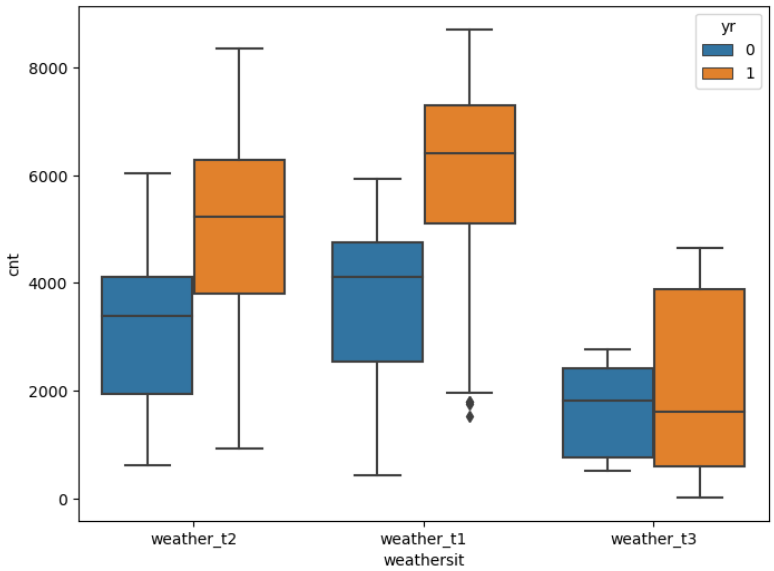
As per observation from distribution with respect to target variable, ‘season’ category shows as follows:



- From the distribution we can observe that many bike bookings are happened during fall season.
- And during spring less number of bookings happened, in year 2018 its much lesser compared with year 2019.
-

(7) weathersit:

As per observation from distribution with respect to target variable, ‘weathersit’ category shows as follows:



The variable weathersit has 4 types of weather conditions as follows.

weather_t1: Clear, Few clouds, Partly cloudy, Partly cloudy

weather_t2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

weather_t3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

weather_t4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

But from the data we have observed that there is no data available on weather type – 4

and from the observation from distribution:

- Less number of bookings happened where there is ab-normal weather conditions like weather_t3.
- Most bookings are happened when there is a clear weather.

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

Answer:
In general dummy variable are created for independent, unordered relative categorical variables. So creating dummy variable for all categories will lead to multicollinearity between each individual categories.
Also if there are ‘n’ categories in variable ‘n-1’ categories will represent/explain/pridict nth category.

For example in the case of categorical variable season:

Fall	Winter	Summer	Spring
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

In this case if season dummy values for all winter, summer, spring are ‘0’ this represents season ‘Fall’. So we don’t need to explicitly define the dummy variable for Fall. As fallows.

0	0	0	→ fall
1	0	0	→ winter
0	1	0	→ summer
0	0	1	→ spring

Question 3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

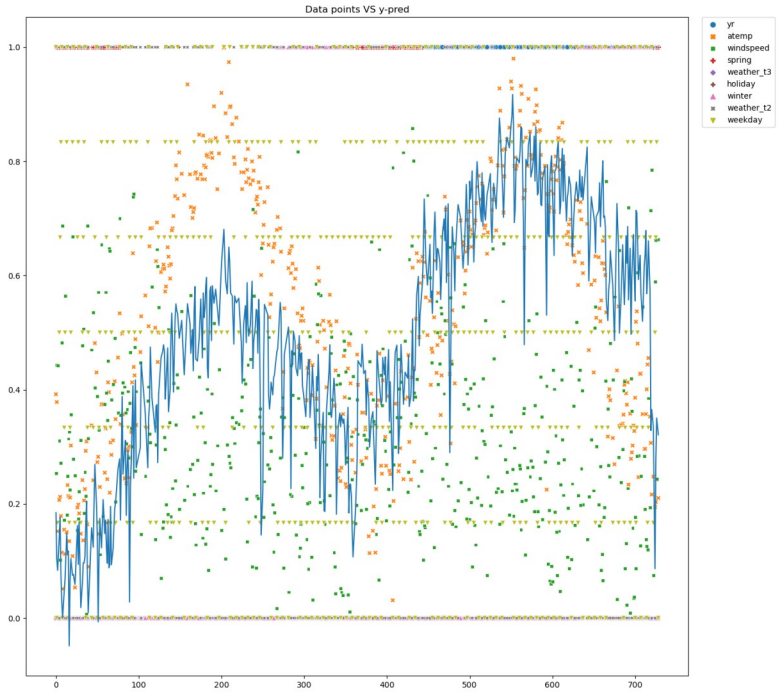
Answer:
The numeric continuous variable **atemp** (feeling temperature)has the highest correlation with target variable.
From the pair plot both temp and atemp have almost same correlation but the correlation value for ‘atemp’ is **0.631059** more compared with ‘temp’.

Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:
There are 3 steps to validate assumptions of linear regression after building model with training set.

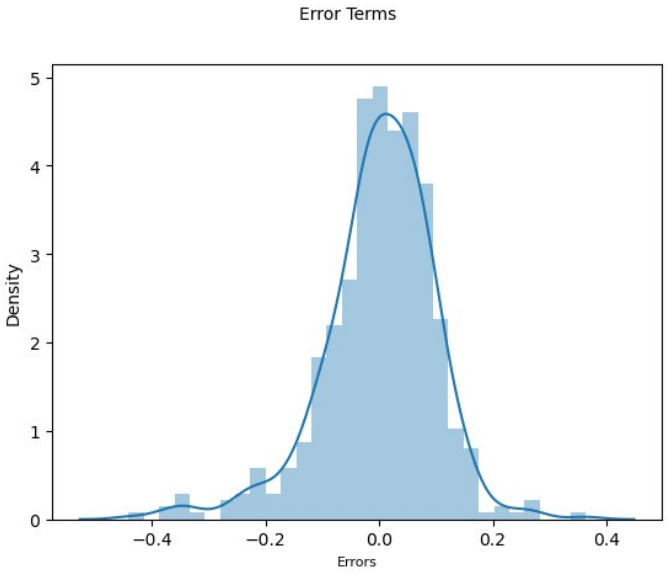
1) **Linear relation ship between independent variables and predicted dependent variable**. By visualizing independent variables along with predicted dependent variable on training will be used for

validating this assumption. For the BikeSharing assignment this assumption is validated using following figure.

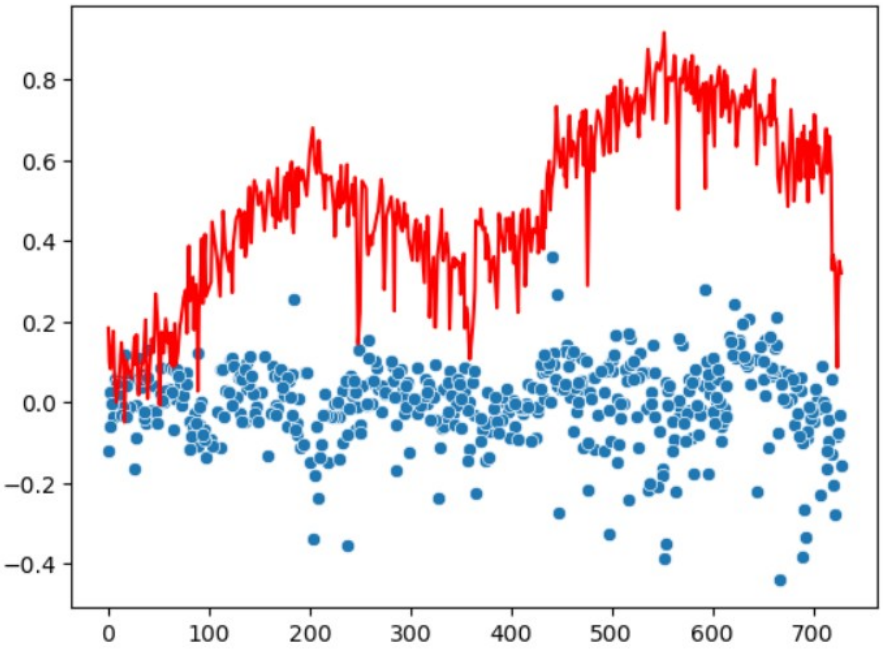


From the above figure we can observe that predicted dependent variable (blue line) and independent variables data points, there is a significant linear relation between independent and dependent target variable from train set.

2) **Residual/error terms distribution follow normal** and centered at 0 (mean=0). By visualizing distribution of error terms we have validated this assumption as shown in figure below.



3) **The error terms should not be dependent on one another.** We can validate this by finding residuals on train plotting both predicted dependent variable and residuals as follows.




```
Correlation between residuals and y_train_pred:  0.0
```

From the above plot red=predictions on train set, blue= residuals , we can observe that there is no correlation between them. Also we can see the correlation value between then is '0'.

4) **Error terms have constant variance (homoscedasticity).**
To validate this assumption we can observe the residual scatter plot, if most of the residuals are ranging at -0.2 to 0.2. And when we compute the variance of residuals we got value 0.009 this is shows that all residuals are approximately same.

```
# Assumption 3: Error terms have constant variance:
print("Varience of error terms :", np.var((y_train - y_train_pred)))
```

```
Varience of error terms : 0.009907198510584716
```

Question 5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:

To know which of the feature are significantly contributing to predict dependent variable we need to observe regression coefficients. In our care the coefficient are as follows.

	coef	std err	t	P
const	0.1811	0.026	6.841	0
yr	0.2436	0.009	27.016	0
atemp	0.4664	0.031	14.985	0
windspeed	-0.1102	0.027	-4.068	0
spring	-0.1288	0.016	-7.883	0
weather_t3	-0.2627	0.031	-8.423	0
holiday	-0.0594	0.027	-2.163	0
winter	0.0479	0.013	3.639	0
weather_t2	-0.0775	0.009	-8.182	0
weekday	0.0549	0.013	4.113	0

Omnibus:	79.694	Durbin-Wat	
Prob(Omnibus):	0.000	Jarque-Ber	
Skew:	-0.821	Prob(JB):	
Kurtosis:	5.449	Cond. No.	

As per linear regression equation we can write predicted dependent variable as follows:

y_pred = 0.1811 + (yr * 0.1811) + (atemp * 0.4664) – (windspeed * 0.1102) – (spring * 0.1288)
- (weahter_t3 * 0.2627) - (holiday * 0.0594) + (winter * 0.0479)
- (weather_t2 * 0.0775) + (weekday * 0.0549)

So from the above equation we can list top tree features which are significantly explains the demand of shared bikes are as follows.

1. **atemp** (positive demand
2. weather_t3(**Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds**) (negative demand)
3. **yr** (positive demand)

Here yr is an independent catagorical variable so we can also consider next highest contributing feature **season-spring**(negative demand)

General Subjective Questions:

Question 1: Explain the linear regression algorithm in detail. (4 marks)

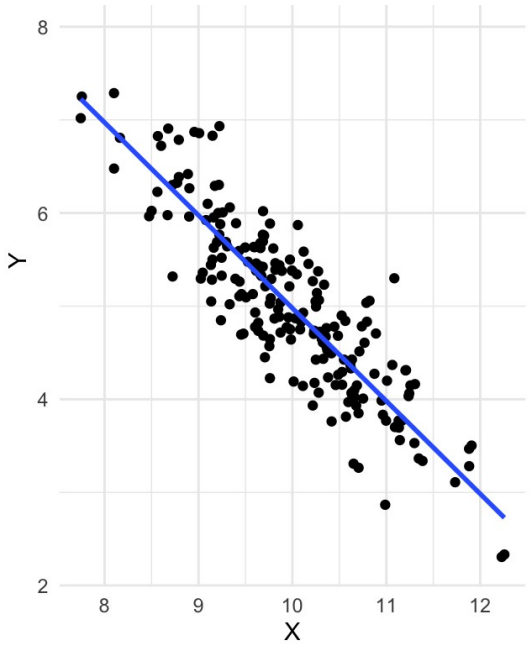
Linear regression is a supervised machine learning algorithm. This algorithm is built based on a simple line formula. This algorithm mainly used for predicting continuous values using one or many independent features or variables by building a statistical model between variables. This algorithm assumes linear relation between input independent and output dependent variables so it is named as linear regression.

Linear regression can be used when we have a dataset with continuous variables and trying to predict a continuous dependent variable using multiple independent variables.

A simple linear regression with one independent variable can be described as follows:

$$Y = \beta 0 + \beta 1 * X + \varepsilon$$

Where Y= Predicted dependent variable
X= input independent variable.
β1 = slope of the regression line.
β0 = intercept (base value of Y when X=0).
ε = error term.
Following figure describes the simple linear regression model.



The main objective of the linear regression is to find the optimal fitting line that minimizes the squared error terms between actual expected observed values and predicted Y(Y_{pred}) values. This wll be done by minimizing **residual sum of squares** or the **mean square error**.

From the equation above we know that the X value is the input and available for predicting Y value, but to estimate Y value we need additional parameters β0, β1. These parameters can be estimated using Ordinary least squares ,method by minimizing the sum of squared errors.

Linear regression has some assumptions to be satisfied in order to achieve best fit model as follows.

- The residuals (actual Y – predicted Y) for all data points should be normally distributed with center 0, mean value(0)
- The independent variables have the linear correlation with dependent variable/predicted Y values.
- The variance of the each residual at each data point should be constant this is call as Homoscedasticity.
- Observations/dependent variable values are independent of each other.

The above shown equation is used when there is a one independent variable, but in realtime application there will be more number of independent variables used to estimate dependent variable. Process of predicting Y value with multiple independent variables using linear regression is also called as multiple linear regression. And the representation of dependent variable is as follows.

$$Y = \beta0 + \beta1*X1 + \beta2*X2 + ... + \beta n*Xn + \varepsilon$$

Where β1, β2 ….. βn denotes regression coefficients for independent variables X1,X2…..Xn

In case of multiple linear regression model need to optimize n number of coefficients using ordinary least squares method. This will be more complex process to achive using normal matrix operations. So to simplify the process of optimization we use **gradient descent** method.
The functionality of the gradient descent method can be described using following equation.

$$W_i = W_{i-1} - \alpha \cdot \nabla J(W_{i-1})$$

Where W_i is the updated coefficient/parameter

W_{i-1} is the previous parameters

α is learning rate

$\nabla J(W_{i-1})$ gradient of the objective function.

parameters will update in the opposite direction of the gradient to minimize the objective function

Once the parameters are optimized and model has been build the next step is to evaluate the model. We can interpret the coefficients (βn) to understand the relationship between the independent and dependent variables.

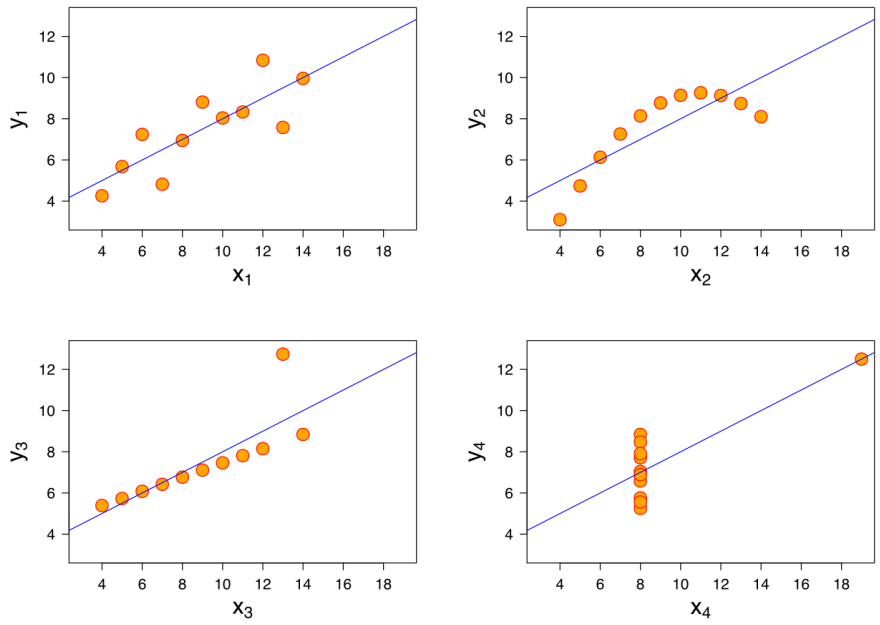
Following factors can be used to evaluate the model:

- R-squared : Denotes proportion of the variance in the dependent variable that is predictable from the independent variables.
- Mean squared error (MSE): Measures the average of the squares of the errors terms.
- Root mean squared error (RMSE): The square root of the MSE, which provides an interpretable measure of the average magnitude of the residuals.

Question 2: Explain the Anscombe's quartet in detail. (3 marks)

Answer:
Anscombe's quartet comprises four datasets with nearly same statistics. This shows the limitations of relaying only on numerical metrics like mean, variance, R-squared and correlations even linear regression lines numerically. But they are very different in distribution wise we can observe visually. This will shows the importance of graphical representation of datasets instead only dependence on numbers.

Following figures shows the example for anscombe's quartet with linear regression and data.



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

Note: Images shown above sourced from google.

Data1: This dataset have linear relationship between X, Y without any outliers Top Left.
Data2: This dataset have no linear relation between X,Y as shown figure Top Right. However it follows normal distribution.
Data3: This dataset follows linear relation between X,Y with one influential outlier Bottom Left.
Data4: This dataset does not follow linear relation or simple relation between X,Y. It has three clusters with similar means, variance and correlations Bottom right.

When we visually observe data it shows different patterns but if we compute statistics of same data results are as follows.

Data1:

- mean of X=9
- Standard deviation of X = 3.16
- Mean of Y = 7.5
- Standard deviation of Y = 1.94
- Correlation X and Y = 0.816

Data2:

- mean of X=9
- Standard deviation of X = 3.16
- Mean of Y = 7.5
- Standard deviation of Y = 1.94
- Correlation X and Y = 0.816

Data3:

- mean of X=9
- Standard deviation of X = 3.16
- Mean of Y = 7.5
- Standard deviation of Y = 1.94
- Correlation X and Y = 0.816

Data4:

- mean of X=9
- Standard deviation of X = 3.16
- Mean of Y = 7.5
- Standard deviation of Y = 1.94
- Correlation X and Y = 0.816

So Anscombe’s quartet shows the importance of visualization of data is for understanding the structure of data instead concluding only using identical summary statistics. This highlighting the danger of making conclusions based on summary statistics.

Question 3: What is Pearson's R? (3 marks)

Answer:
Pearson's R is a correlation coefficient, this measures strength of linear correlation between two sets of data. The pearson's R correlation computed from ratio between covariance of two variables and the product of their standard deviations. This method introduced by Karl Pearson that is why named as Pearson's R where R mean correlation coefficient.

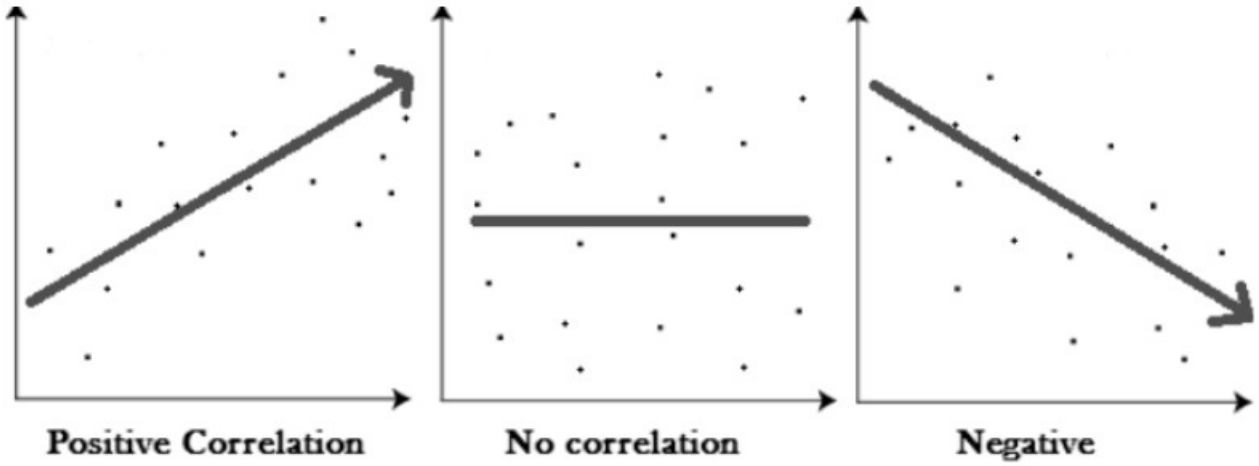
Following equation describes pearson's correlation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where n is number of data points

x_i, y_i are individual data points in the two variables X,Y.

The Pearson's correlation coefficient ranges from -1 to 1. where sign indicates negative or positive correlation. And value indicates how perfectly the sets are correlated with each other.
Following figure shows the example show pearson's correlation look like.



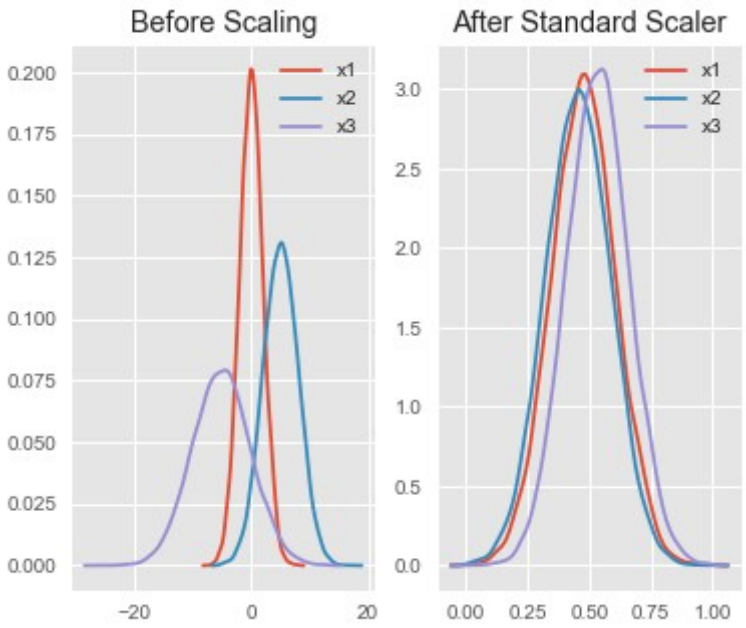
- Positive correlation: As one increases other increases.
- Negative correlation : As one increases other decreases

Pearson's correlation assumes input sets are linearly correlated with each other, so it will show how well the variable data points are linearly correlated with each other. So this method not captures nonlinear relationships. This method is very sensitive to outliers. This method is mostly used in predictive tasks due to its interpretability.

Question 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Marks 3)

Sailing is the process of transforming the value ranges of features or dependent variables. Scaling generally used to make variables contribution equally for analysis. This help avoiding model biasing towards larger scale values. This will also help comparing and visualization of multi range features in to scaled dimension.

As per research scaling is very impotent in gradient descent optimization, because features with high scales will influence the results. The following figure shows example of scaling values.



There are several types of scaling methods used in machine learning algorithms, most commonly used once are normalized scaling , Standardized scaling the difference as follows.

Normalized scaling(Min-Max scaling)
- The range is fixed between 0 to 1
- Formula for computing scaling :

$$x_{\text{scaled}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Useful when distribution of the data is uniform or unknown.
- Non-sensitive to outliers because data ranges from 0 to 1
- Computed from minimum and maximum values

Standardized scaling(Z-score normalization)
- No fixed range transforms data to have
- Formula for computing scaling :

$$x_{\text{scaled}} = \frac{x - \text{mean}(x)}{\text{std}(x)}$$

Where std(x) is the standard deviation of x

- Useful when distribution of the data is normal
- Sensitive to outliers because there is no fixed range.
- Fits data to get mean as 0 and standard deviation 1

Question 5: You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

VIF stands for Variance inflection faction, used for computing multicollinearity in regression analysis. Multicollinearity occurs when there is a high correlation between input data points or independent variables. High VIF affects interpretation of coefficients.

In general we can say the VIF is how much variance explained by other independent variables. This can be measured using following formula.

$$\mathrm{VIF}_i = \frac{1}{1-R_i^2}$$

Where VIF_i is the Variance inflection factor for i_{th} variable.

R_i^2 Is the unadjusted coefficient of determination computed by regression for ith independent variable.

If VIF = 1, means there is no multicollinearity in between input independent variables.

If VIF >1 and VIF<5: Moderate multicollinearity between input independent variables with current variable. Means approximately below 80% of variance is explained by other independent variables.

If VIF >5 High multicollinearity between current and input independent variables. And more then 80% of variance is explained by other variables. This range is problematic in model development and we should definitely drop the variable from the analysis.

If VIF if infinite this means there is one or more independent variables are **100% matching** with each other or might be **duplicate** data with different variable name. For example there are 2 independent variables shown in figure.

X1	X2	X3
2	2	1
6	6	3
7	7	4
5	5	0
4	4	6

In this case for the first time model development both X1,X2 variables shows VIF as infinite.
If we drop any one of them from the analysis VIF of other (X1 or X2) will be calculated with X2 and it will not be infinite because they are not exactly same.

Question 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

Quantile-Quantile (Q-Q) plot is used to assess as set of data points distribution. This plot compares quantiles of data with quantiles of specific distribution. In general this plot compares with normal distribution quantiles.

Steps for generating Q-Q plot:
- Sorting data in ascending order.
- Compute quantiles of data.
- Compute quantiles of corresponding distribution.
- Plotting the quantiles of the data against quantiles of corresponding distribution.

Impotence of Q-Q plots:

- This plots are commonly used to validate the assumptions. For example in case of linear regression assumption validation like normal distribution we can use Q-Q plot. If the residuals are deviated significantly from straight line on the plot we can conclude that this assumption is not satisfied.

- Q-Q plots are also helps in finding the outliers in data. In Q-Q plot outliers are significantly deviate from the straight line. This will help in making robust models.

- In model validation Q-Q plots are used as testing tool, and based on plot result we may re-construct and build models.