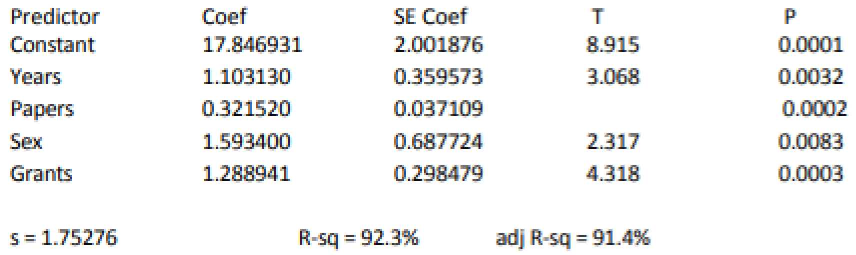
Vivekanand Education Society's Institute of Technology, Chembur, Mumbai,
Department Of Artificial Intelligence and Data Science
Year:2023-24 (Even Sem)
MID TERM TEST

Class : D11AD	**Division: -**
Semester: VI	**Subject: Data Analytics & Visualization**
Date: 26th Feb 2024	**Time: 9 am to 10 am**

Q1. a. Enlist the key elements in a Chart

Data Points, Lines, Bars, Axes, Labels, Title, Legends, Gridlines, Markers, Area Fill, Annotations.

Q1 b. A scientific foundation wanted to evaluate the relation between y= salary of the researcher (in thousands of dollars), x_1= number of years of experience, x_2= an index of publication quality, x_3=sex (M=1, F=0), and x_4= an index of success in obtaining grant support. A sample of 35 randomly selected researchers was used to fit the multiple regression model. Parts of the computer output appear below.



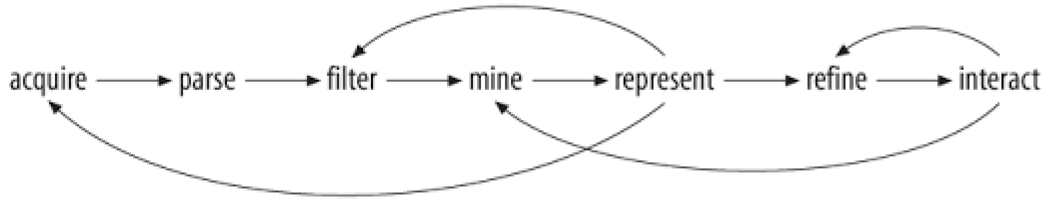
```
Predictor      Coef         SE Coef       T          P
Constant       17.846931    2.001876      8.915      0.0001
Years          1.103130     0.359573      3.068      0.0032
Papers         0.321520     0.037109                 0.0002
Sex            1.593400     0.687724      2.317      0.0083
Grants         1.288941     0.298479      4.318      0.0003

s = 1.75276              R-sq = 92.3%       adj R-sq = 91.4%
```

(a) **The least squares line fitted to the data :**

$$salary = 17.85 + 1.10\,x_1 + 0.32\,x_2 + 1.59\,x_3 + 1.29\,x_4$$

(b) **How many degrees of freedom does the t* value from the previous question have?** 30

Q1 c. With a neat diagram explain briefly the 7 stages involved in Visualizing Data.



1. **Acquire** : Obtain the data, whether from a file on a disk or a source over a network.
2. **Parse** : Provide some structure for the data's meaning, and order it into categories.
3. **Filter :** Remove all but the data of interest.
4. **Mine :** Apply methods from statistics or data mining as a way to discern patterns or place the data in mathematical context.
5. **Represent** : Choose a basic visual model, such as a bar graph, list, or tree.
6. **Refine :** Improve the basic representation to make it clearer and more visually engaging.
7. **Interact** : Add methods for manipulating the data or controlling what features are visible.

Q1. d. Enlist the key roles involved in a Data Analytics Project

1. Business User – understands the domain area
2. Project Sponsor – provides requirements

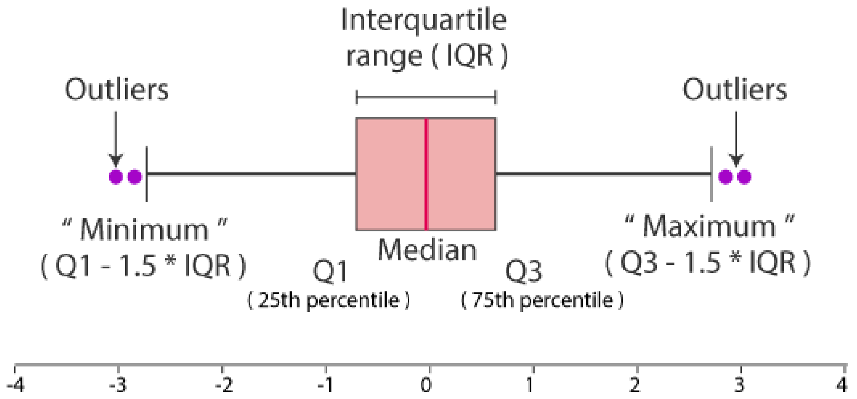
3. Project Manager – ensures objectives
4. Business Intelligence Analyst – provides business domain expertise based on deep understanding of the data
5. Database Administrator (DBA) – creates DB environment
6. Data Engineer – provides technical skills, assists data management and extraction, supports analytic sandbox
7. Data Scientist – provides analytic techniques and modeling

---

Q1. e. Enlist 5 tools used for Narrative Visualization

- **Tableau** is a widely used data visualization and business intelligence tool that allows users to create interactive and shareable dashboards. It supports the creation of compelling narrative visualizations with features like story points, allowing users to guide viewers through a series of data-driven insights.
- **Microsoft Power BI** is another powerful tool for creating narrative visualizations. It enables users to connect to various data sources, create dynamic reports and dashboards, and use storytelling features to guide the audience through the data-driven narrative.
- **D3.js (Data-Driven Documents)** is a JavaScript library for creating dynamic and interactive data visualizations in web browsers. While it requires more coding skills compared to other tools, D3.js provides unparalleled flexibility for crafting customized narrative visualizations on the web.
- **RAWGraphs** is an open-source web application that simplifies the creation of custom visualizations. It provides a user-friendly interface for designing various chart types and allows users to create story-driven visualizations by importing their data and customizing the appearance of the charts.
- **StoryMapJS** is a web-based tool for creating interactive, map-based narrative visualizations. It is particularly useful for projects that involve geographical data or timelines. Users can create slides with images, text, and multimedia elements to guide the audience through a story tied to geographic locations.

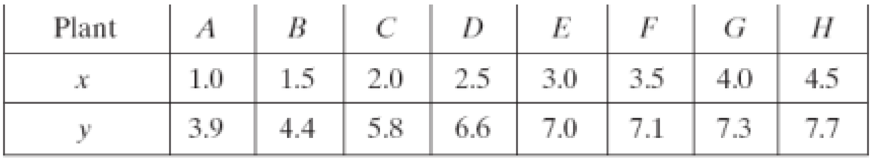
Q1. f. Explain the 5 elements of a Box Plot



1. **Minimum and Maximum Values (Whiskers):** represent the range of the data. The lower whisker extends from the minimum value to the first quartile (Q1), and the upper whisker extends from the third quartile (Q3) to the maximum value. Outliers beyond the whiskers may also be plotted individually.
2. **Box (Interquartile Range - IQR):** represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). The length of the box illustrates the spread of the central 50% of the data. The horizontal line inside the box represents the median (Q2), which is the midpoint of the dataset.
3. **Quartiles (Q1, Q2, Q3):** divide the dataset into four equal parts, each containing approximately 25% of the data.
   - Q1 (First Quartile): The value below which 25% of the data falls.
   - Q2 (Second Quartile/Median): The middle value of the dataset, separating the lower and upper halves.
   - Q3 (Third Quartile): The value below which 75% of the data falls.

4. **Outliers:** Individual data points that lie beyond the whiskers are considered outliers. Outliers can provide information about the variability and potential skewness in the data. They are often plotted as individual points or marked differently for emphasis.
5. **Notches:** on either side of the box, which provide a visual estimate of the uncertainty around the median. If the notches of two box plots do not overlap, it suggests a significant difference in medians at a certain level of confidence.

Q2. b. A biologist assumes a linear relationship between the amount of fertilizer supplied to tomato plants and the subsequent yield of tomatoes obtained. He randomly selected 8 tomato plants of the same variety and treated them weekly with a solution in which x grams of fertilizer was dissolved in a fixed quantity of water. The yield in y kilograms of tomatoes was recorded.

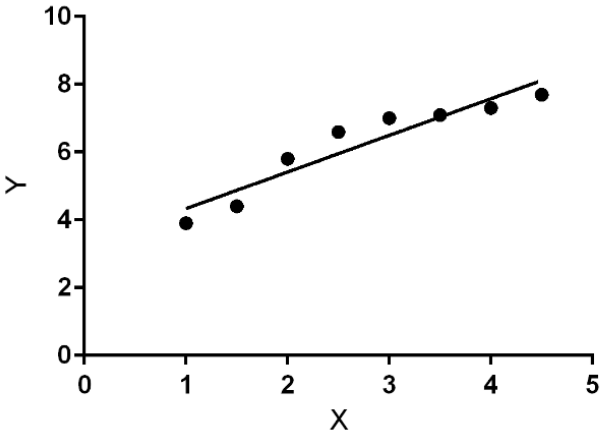


Plant	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
x	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
y	3.9	4.4	5.8	6.6	7.0	7.1	7.3	7.7

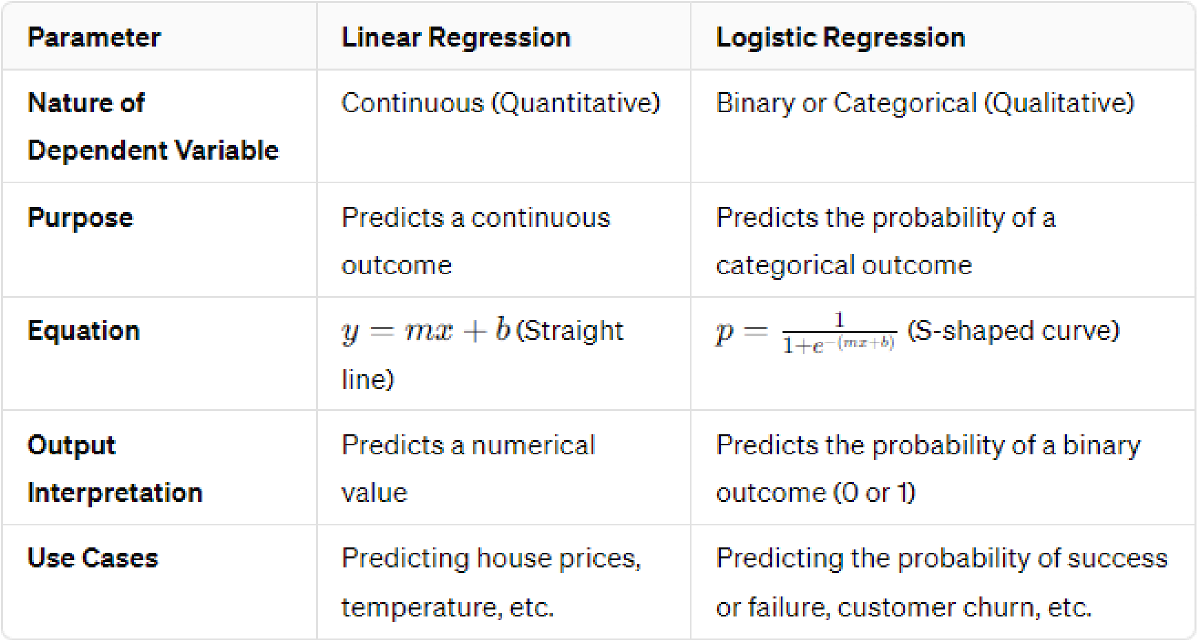
(a) **Calculate the equation of the Least Squares Regression line of y on x. :**

Y = 1.081*X + 3.252

(b) **Estimate the yield of a plant treated, weekly with 3.2 grams of fertilizer.** 6.7112

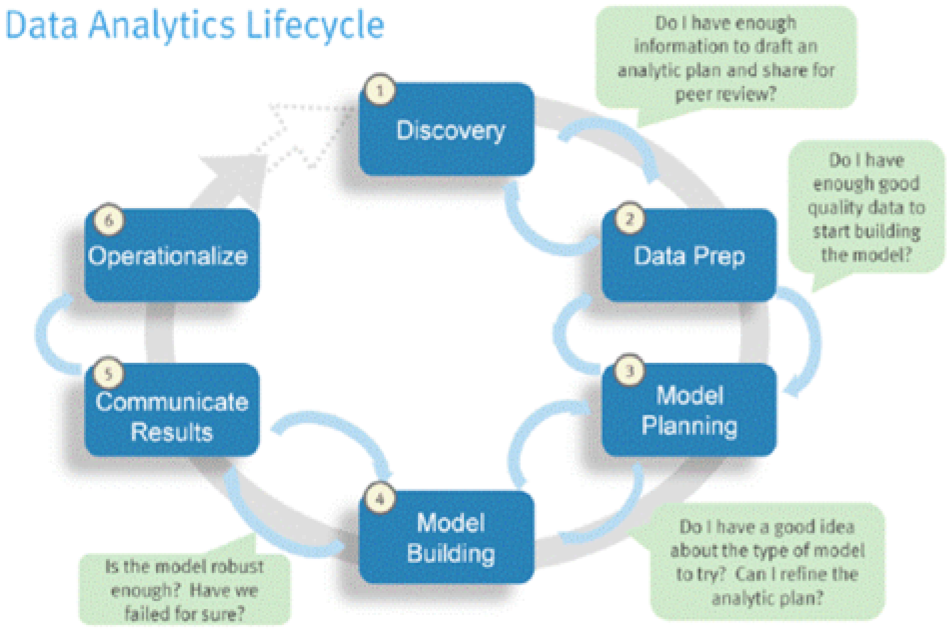


Q3. a. Compare Linear Regression with Logistic Regression with 5 parameters



Parameter	Linear Regression	Logistic Regression
Nature of Dependent Variable	Continuous (Quantitative)	Binary or Categorical (Qualitative)
Purpose	Predicts a continuous outcome	Predicts the probability of a categorical outcome
Equation	$y = mx + b$ (Straight line)	$p = \frac{1}{1+e^{-(mx+b)}}$ (S-shaped curve)
Output Interpretation	Predicts a numerical value	Predicts the probability of a binary outcome (0 or 1)
Use Cases	Predicting house prices, temperature, etc.	Predicting the probability of success or failure, customer churn, etc.

Q3. b. With a neat diagram explain the Data Analytics Lifecycle. For each phase write down the steps involved.


Data Analytics Lifecycle
1
Discovery
Do I have enough information to draft an analytic plan and share for peer review?
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
5
Communicate Results
Is the model robust enough? Have we failed for sure?
6
Operationalize

Phase 1 - Discovery

1. Learning the Business Domain
2. Resources - Available Time, People, Tech, Data
3. Framing the Problem
4. Identifying Key Stakeholders
5. Interviewing the Analytics Sponsor
6. Developing Initial Hypotheses
7. Identifying Potential Data Sources

Phase 2 - Data Preparation

1. Preparing the Analytic Sandbox
2. Performing ETLT
3. Learning about the Data
4. Data Conditioning
5. Survey and Visualize
6. Common Tools for Data Preparation

Phase 3 - Model Planning

1. Data Exploration & Variable Selection
2. Model Selection
3. Common Tools for Model Planning Phase

Phase 4 - Model Building

- Execute the models defined in Phase 3
- Develop datasets for training, testing, and production
- Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.
- Free or open-source tools : Rand PL/R, Octave, WEKA, Python
- Commercial tools : Matlab , STASTICA, SPSS Modeler

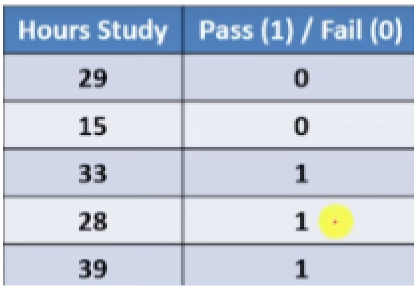
Phase 5 - Communicate Results

- Determine Success / Failure of the team in its objectives
- Assess if the results are statistically significant and valid
- Communicate and document the key findings and major insights derived from the analysis

Phase 6 - Operationalize

- The team sets up a pilot project to deploy the work in a controlled way
- Risk is managed effectively by undertaking small scope, pilot deployment before a wide-scale rollout
- During the pilot project, team executes the algorithm more efficiently in the database
- To test the model in a live setting, consider running the model in a production environment for a discrete set of products or a single line of business
- Monitor model accuracy and retrain the model if necessary

Q2. a. Given a dataset of pass / fail in an exam of 5 students in the given table. Use Logistic Regression as a classifier to answer the following questions. Given the log(odds) = -64 + 2 * hours



Hours Study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
39	1

(a) Calculate the probability of a pass for the student who studied for 40 hours.

(b) At least how many hours does the student study that helps him /her to pass the course with a probability of more than 80%