

DATA SCIENCE BOOTCAMP

MODULE 3

Module Outline

- Data Pre-processing.
- The Use Of Plots.
- Azure ML Hands on Lab.
- Few Questions from the lab.

Data Preprocessing

Why do we need to process the data?



Microsoft



- Data in the real world is dirty
 - **Incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - **Noisy**: containing errors or outliers
 - **Inconsistent**: containing discrepancies in codes or names
- No quality data, no quality mining results!
 - Quality decisions must be based on quality data
 - Data warehouse needs consistent integration of quality data

Measure of Data Quality

- Accuracy
- Completeness
- Consistency
- Timeliness
- Value added
- Interpretability
- Accessibility

Tasks in Data Preprocessing



Microsoft



- Data cleaning
 - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
 - Integration of multiple databases, data cubes, or files
- Data transformation
 - Normalization and aggregation
- Data reduction
 - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
 - Part of data reduction but with particular importance, especially for numerical data

Data Cleaning

- Data cleaning tasks
 - Fill in missing values
 - Identify outliers and smooth out noisy data
 - Correct inconsistent data

Missing Data

- Data is not always available
 - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
 - equipment malfunction
 - inconsistent with other recorded data and thus deleted
 - data not entered due to misunderstanding
 - certain data may not be considered important at the time of entry
 - not register history or changes of the data
- Missing data may need to be inferred.

How to Handle Missing Data?



Microsoft



- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably).
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!
- Use the attribute mean to fill in the missing value.

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
 - faulty data collection instruments
 - data entry problems
 - data transmission problems
 - technology limitation
 - inconsistency in naming convention
- Other data problems which requires data cleaning
 - duplicate records
 - incomplete data
 - inconsistent data

How to Handle Noisy Data?



Microsoft



- Clustering
 - detect and remove outliers
- Combined computer and human inspection
 - detect suspicious values and check by human

Redundant Data

- Redundant data occur often when integration of multiple databases
 - The same attribute may have different names in different databases.
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Integration

- Data integration:
 - combines data from multiple sources into a coherent store
- Detecting and resolving data value conflicts
 - for the same real world entity, attribute values from different sources are different
 - possible reasons: different representations, different scales, e.g., metric vs. British units

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Normalization: scaled to fall within a small, specified range
 - min-max normalization
 - z-score normalization
 - normalization by decimal scaling

Data Transformation: Normalization

- min-max normalization

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- z-score normalization

$$v' = \frac{v - \text{mean}_A}{\text{stand_dev}_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$



- Feature selection (i.e., attribute subset selection):
 - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
 - reduce # of patterns in the patterns, easier to understand

Subset Selection

- Two approaches
 - Forward Selection
 - Backward Selection
- Forward Selection: Start of with 0 features and add them until the error rate for a given evaluator decreases.
- Backward Selection: Start of with N features and reduce them until the error rate for a given evaluator decreases.

Exploratory Data Analysis

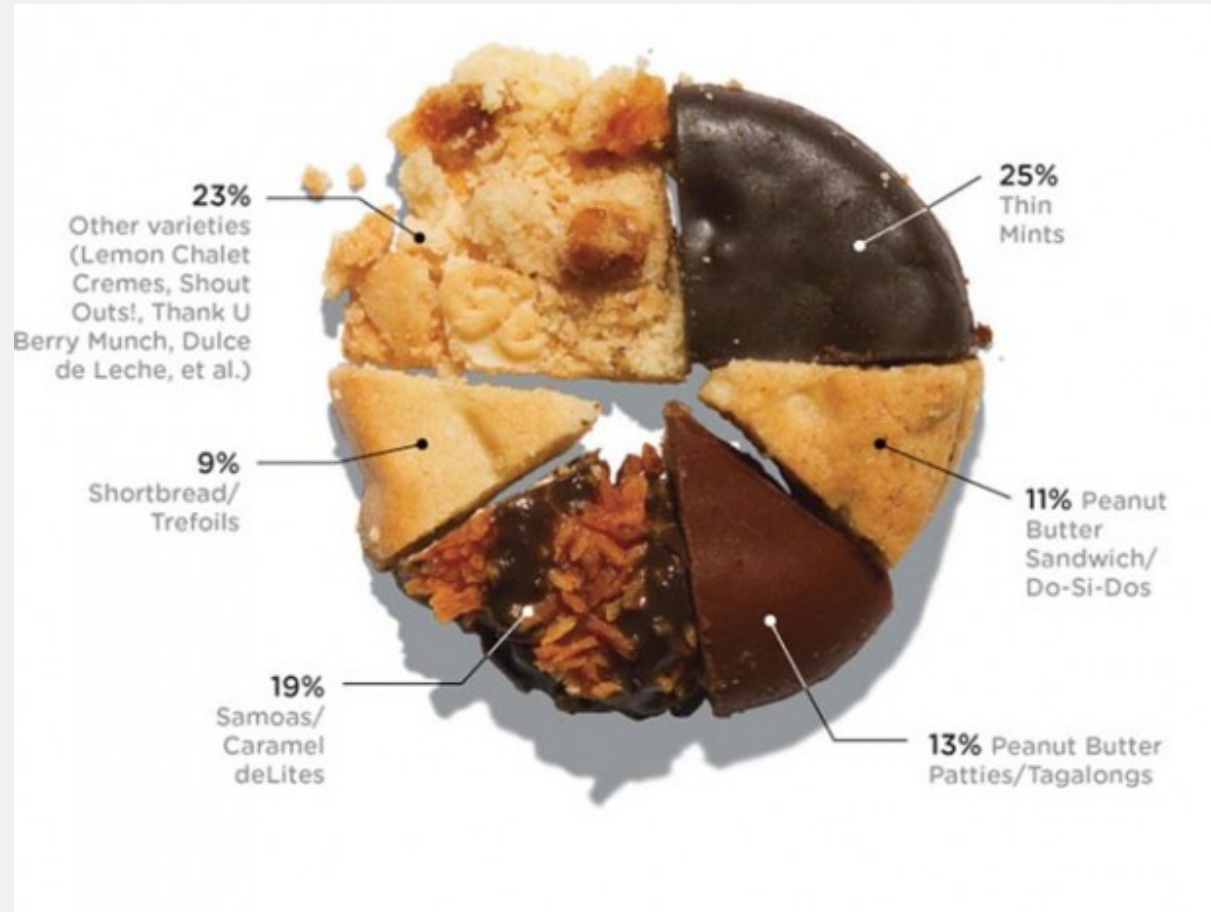


- Exploratory Data Analysis (EDA) and Visualization are very important steps in any analysis task.
- Get to know your data!
 - distributions (symmetric, normal, skewed)
 - data quality problems
 - outliers
 - correlations and inter-relationships
 - subsets of interest
 - suggest functional relationships



- Goal: get a general sense of the data
 - means, medians, quantiles, histograms, boxplots
 - You should always look at every variable - you will learn something!
- Think interactive and visual
 - Humans are the best pattern recognizers
 - You can use more than 2 dimensions!
 - x,y,z, space, color, time, etc.
- Especially useful in early stages of data mining
 - detect outliers (e.g. assess data quality)
 - test assumptions (e.g. normal distributions or skewed?)
 - identify useful raw data & transforms (e.g. $\log(x)$)

Data Visualization



- Sample statistics of data X
 - mean: $\mu = \sum_i X_i / n$
 - mode: most common value in X
 - median: $X = \text{sort}(X)$, median = $X_{n/2}$ (half below, half above)
 - quartiles of sorted X : Q1 value = $X_{0.25n}$, Q3 value = $X_{0.75n}$
 - interquartile range: value(Q3) - value(Q1)
 - range: $\max(X) - \min(X) = X_n - X_1$
 - variance: $\sigma^2 = \sum_i (X_i - \underline{\mu})^2 / n$

The Use Of Plots

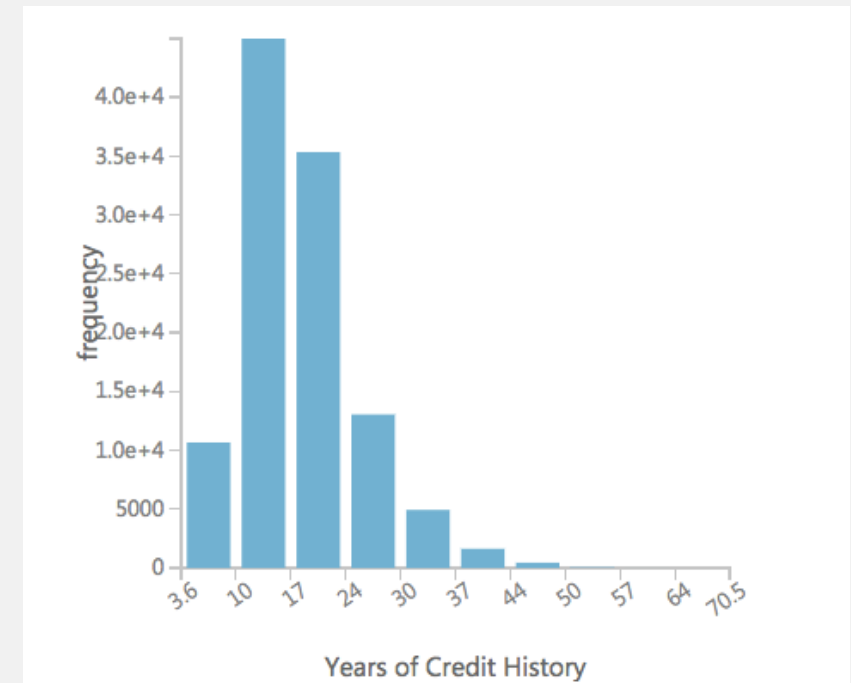
Single Variable Visualization



Microsoft



- Histogram:
 - Shows center, variability, skewness, modality, outliers, or strange patterns.

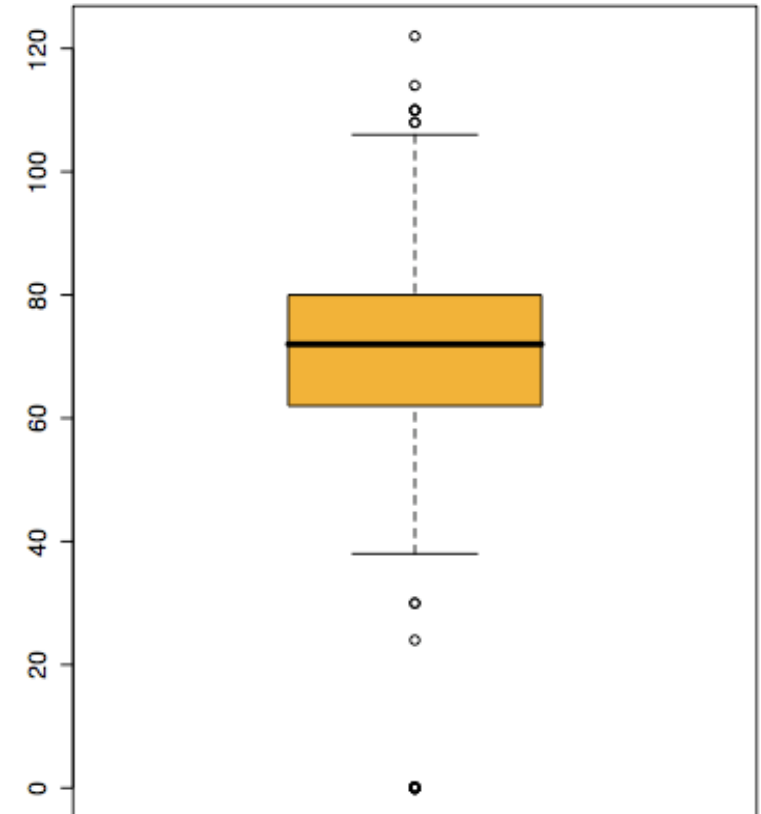


Issues with Histograms

- For small data sets, histograms can be misleading.
 - Small changes in the data, bins, or anchor can deceive
- For large data sets, histograms can be quite effective at illustrating general properties of the distribution.
- Histograms effectively only work with 1 variable at a time
 - But 'small multiples' can be effective

Boxplots

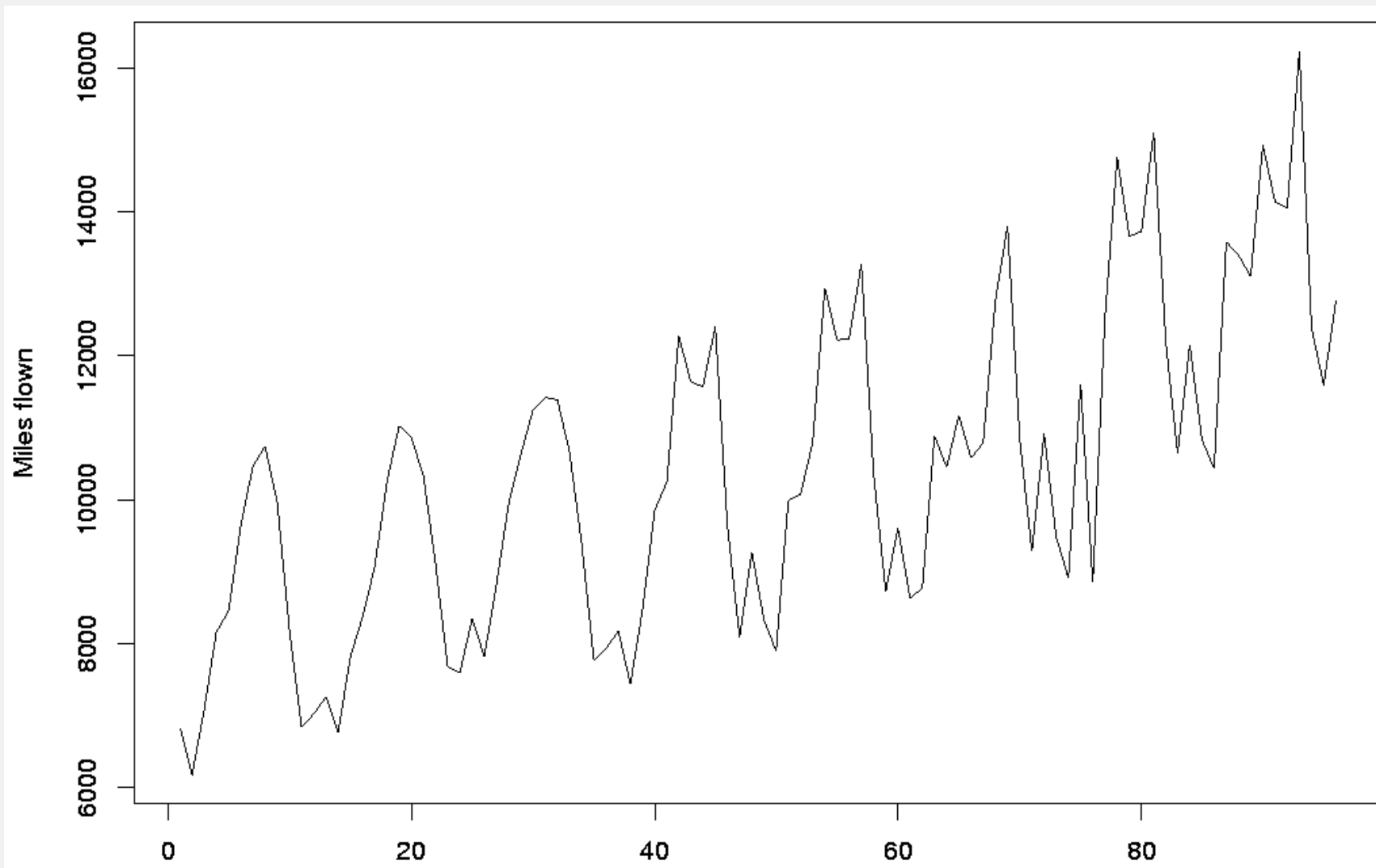
- Shows a lot of information about a variable in one plot
 - Median
 - IQR
 - Outliers
 - Range
 - Skewness
- Negatives
 - Over-plotting
 - Hard to tell distributional shape
 - no standard implementation in software (many options for whiskers, outliers)



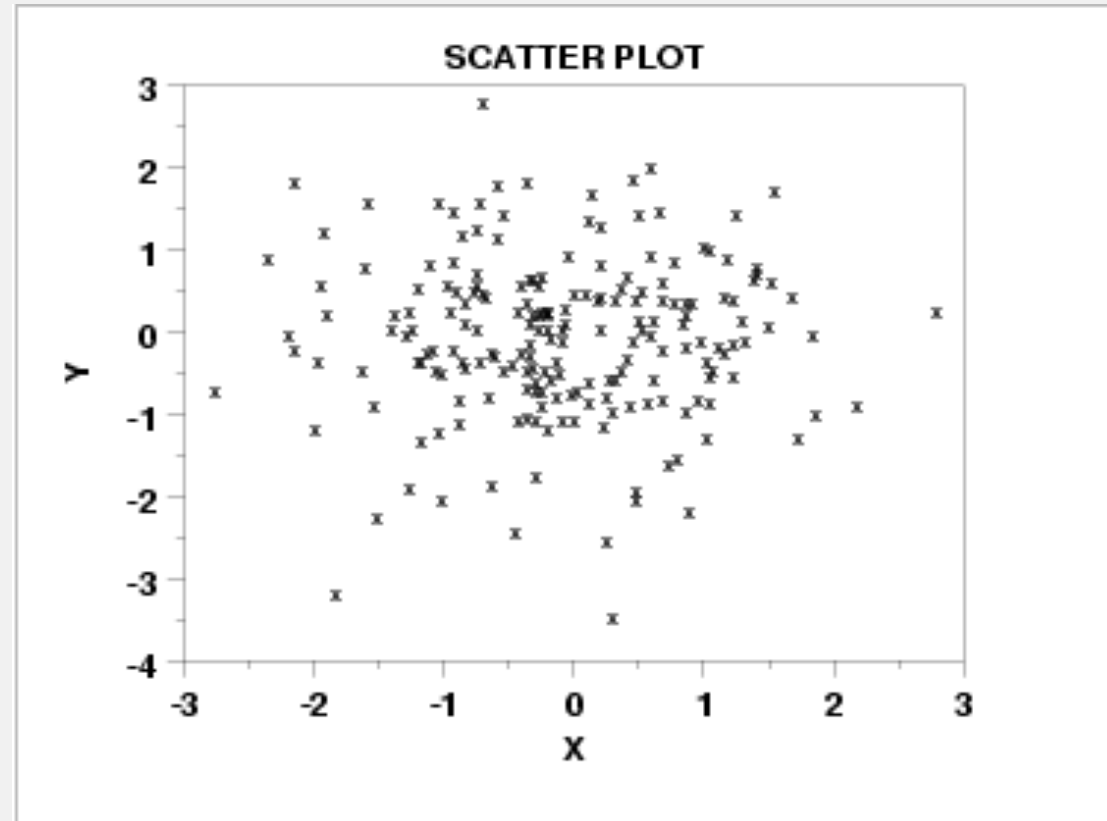
Time Series



Microsoft



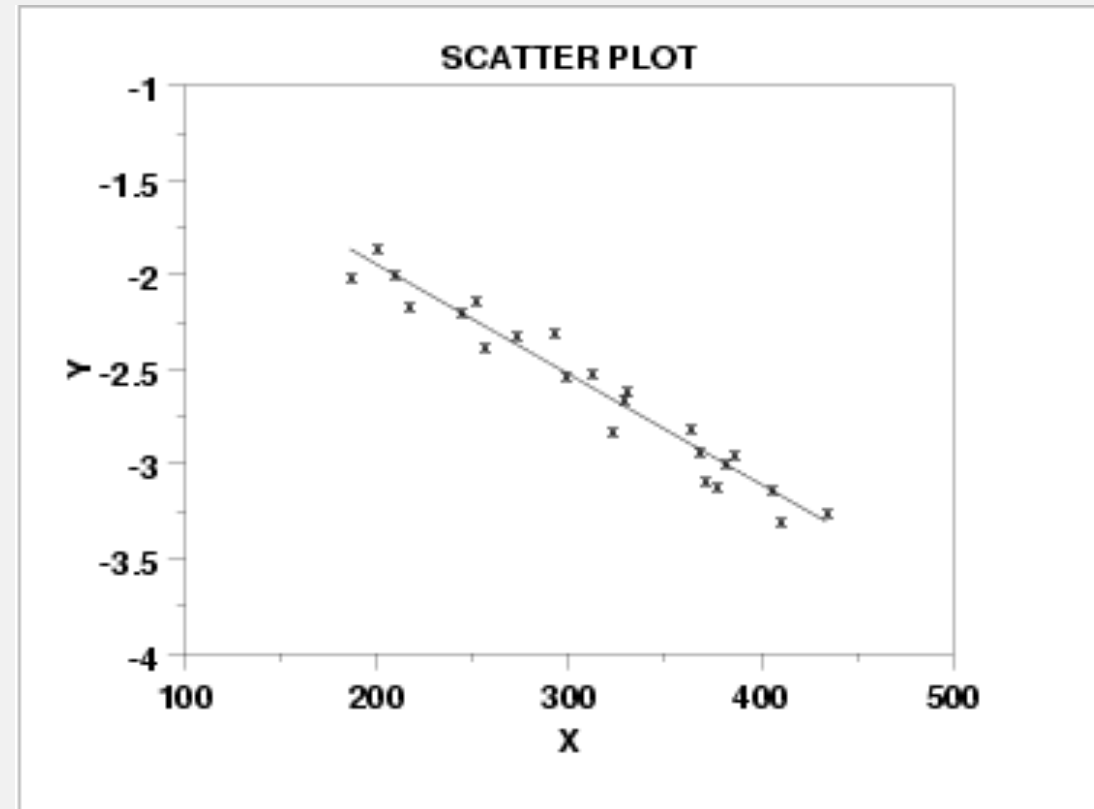
Scatter Plot: No Apparent Relationship



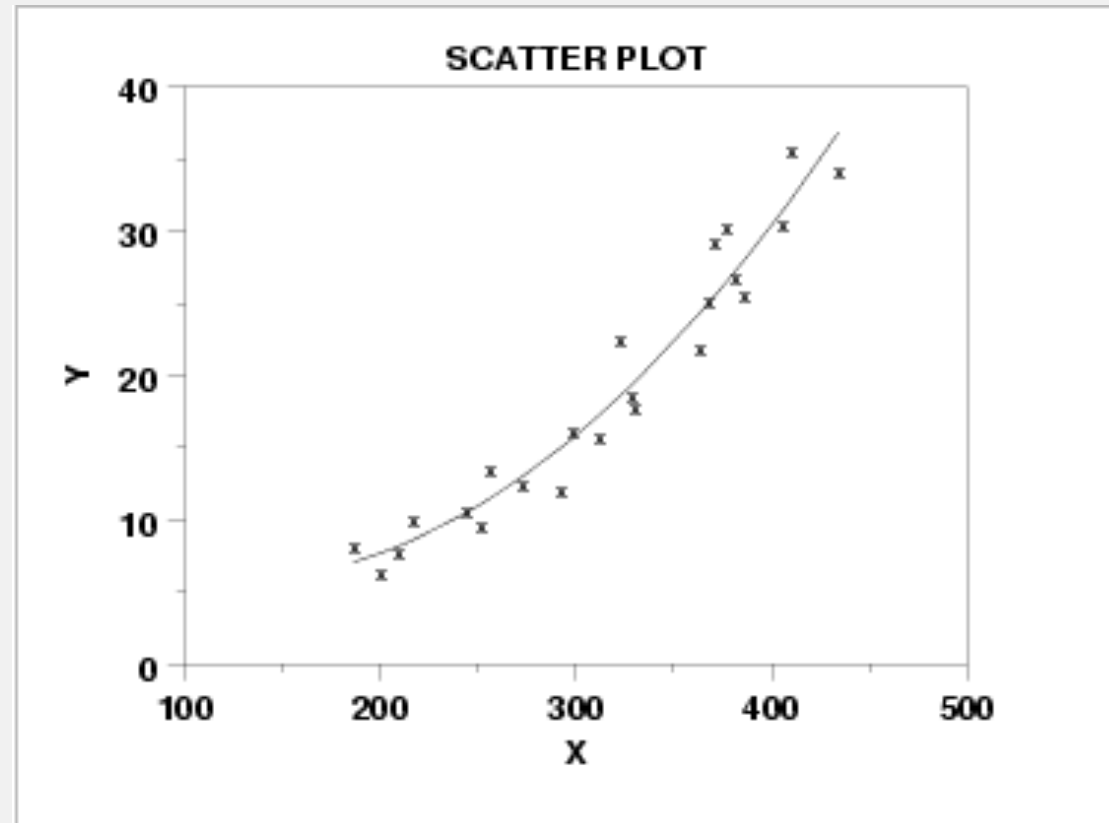
Scatter Plot: Linear Relationship



Microsoft



Scatter Plot: Quadratic Relationship



Scatter Plots

- Scatterplots
 - But can be bad with lots of data

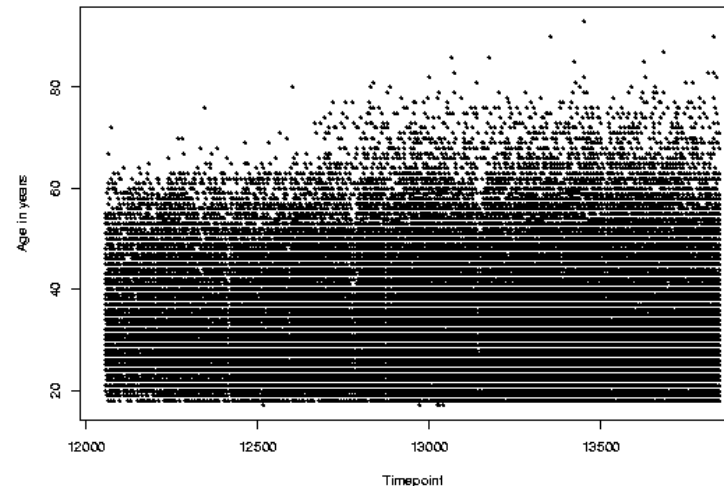
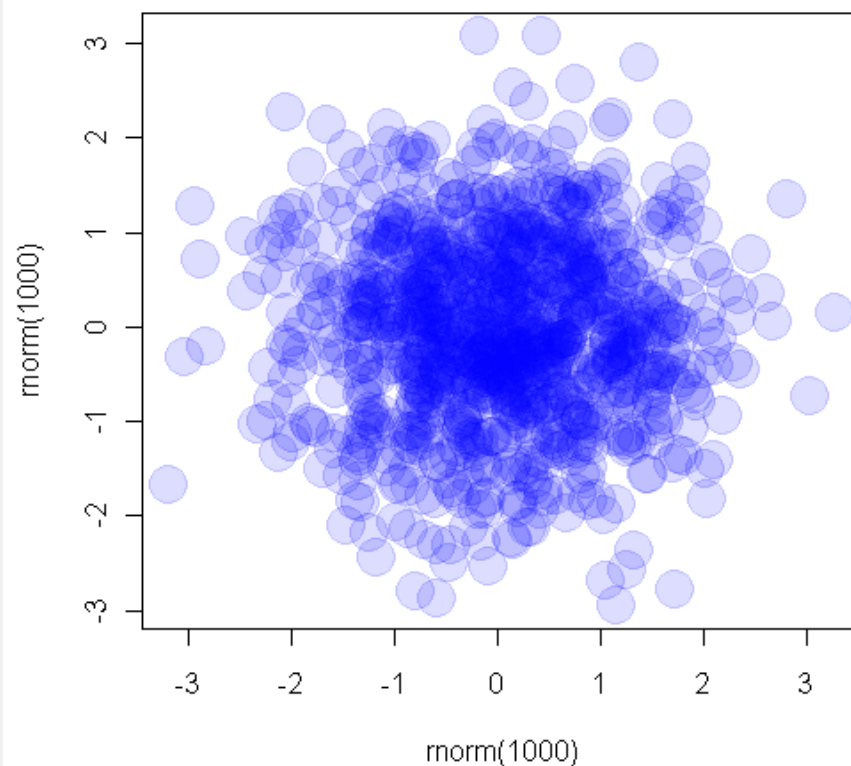


Figure 3.7: A scatterplot of 96,000 cases, with much overprinting. Each data point represents an individual applicant for a loan. The vertical axis shows the age of the applicant, and the horizontal axis indicates the day on which the application was made.

Transparent Plotting



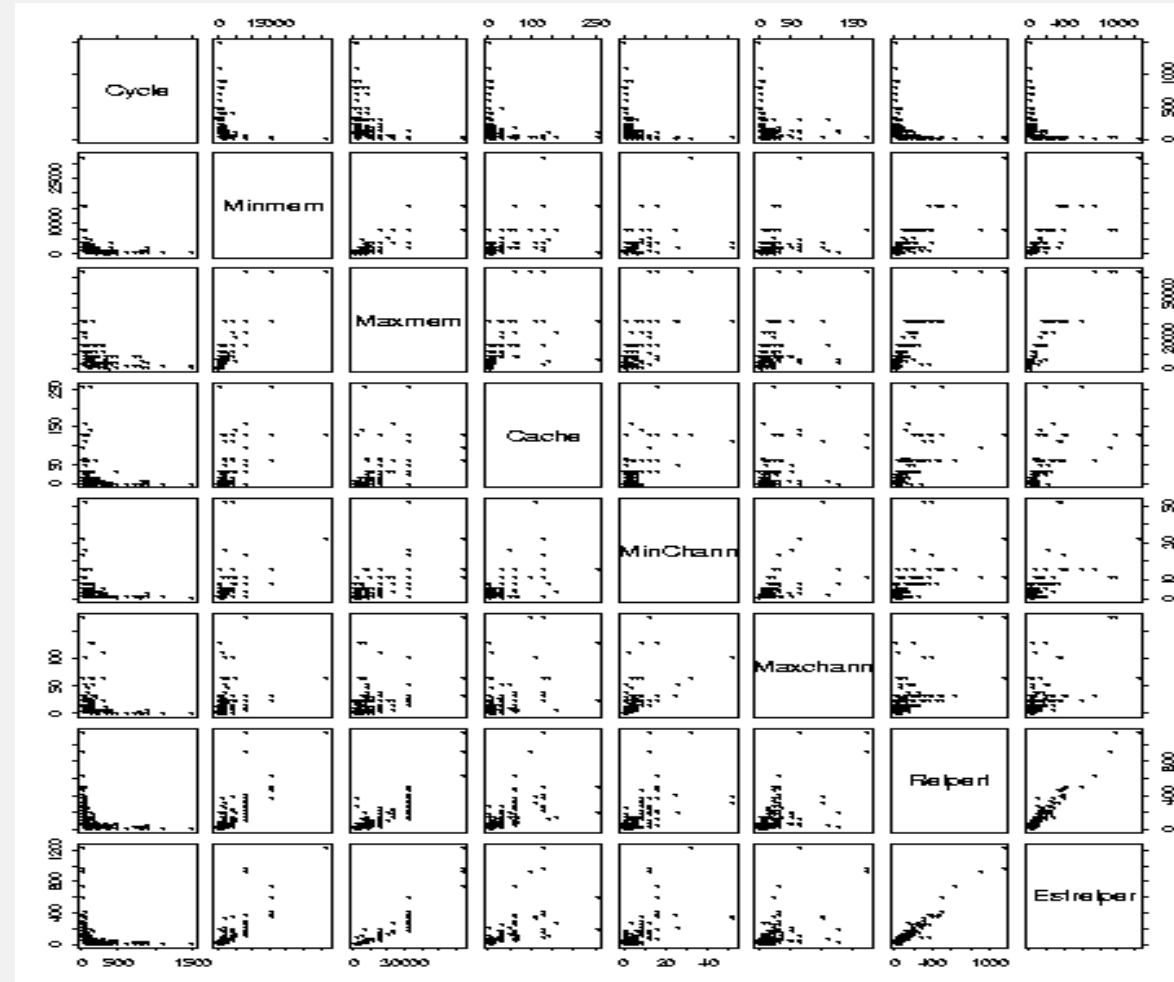
Microsoft



More Than Two Variables



Microsoft



Azure Pass

Home - Microsoft Azure Pass x A ahad

Secure | <https://www.microsoftazurepass.com>

Apps ★ Bookmarks Book Statistics for... business-statistics... Course Outline MT... Preamble to the co... Statistics for Busin... Weiss Introductory... >> Other Bookmarks

Microsoft Azure MY ACCOUNT SIGN IN

Ready to get started?

Try Microsoft Azure Pass

We're offering an Azure Pass, so for a limited time period, you can try Azure for free

*No credit card required

Start >

Use the links below to learn more

- [Redemption Process Guide](#)
- [Azure Documentation](#)
- [Explore Azure](#)
- [GDPR Documentation](#)



Microsoft



Microsoft Azure

[MY ACCOUNT](#)

[SIGN OUT](#)

The following Microsoft Account will be used for Azure Pass:

Given name: Ahad

Surname: Mushir

Microsoft Email: ahadmushir@outlook.com

If the above email address is incorrect, please [sign out](#) and redeem using the correct Microsoft Account

[Confirm Microsoft Account >](#)

Microsoft Azure

[MY ACCOUNT](#)[SIGN OUT](#)

The following Microsoft Account will be used for Azure Pass:

Given name: Ahad

Surname: Mushir

Microsoft Email: ahadmushir@outlook.com

If the above email address is incorrect, please [sign out](#) and redeem using the correct Microsoft Account

Enter Promo code:

[Claim Promo Code](#)

Microsoft Azure

MY ACCOUNT

SIGN OUT

Thank you for redeeming an Azure Pass

This Azure Pass offer provides the following:

- \$50 USD monthly credits (converted to local currency)
- 1 month duration
- View [offer details](#)

Get started now:

Activate >

Use the links below to explore Azure

[Explore Azure](#)

[Azure Documentation](#)



Microsoft



Microsoft Azure

Purchase

ahadmushir@outlook.com | Sign Out

Azure Pass

Learn more ▼

1 Agreement

☒ I agree to the [subscription agreement](#), [offer details](#), and [privacy statement](#).

Purchase →

Azure ML Studio



- Log on to <https://studio.azureml.net/>
- If you don't have an account so browse to sign up **Free Workspace** option with your existing Microsoft account.

Uploading an IPython Notebook

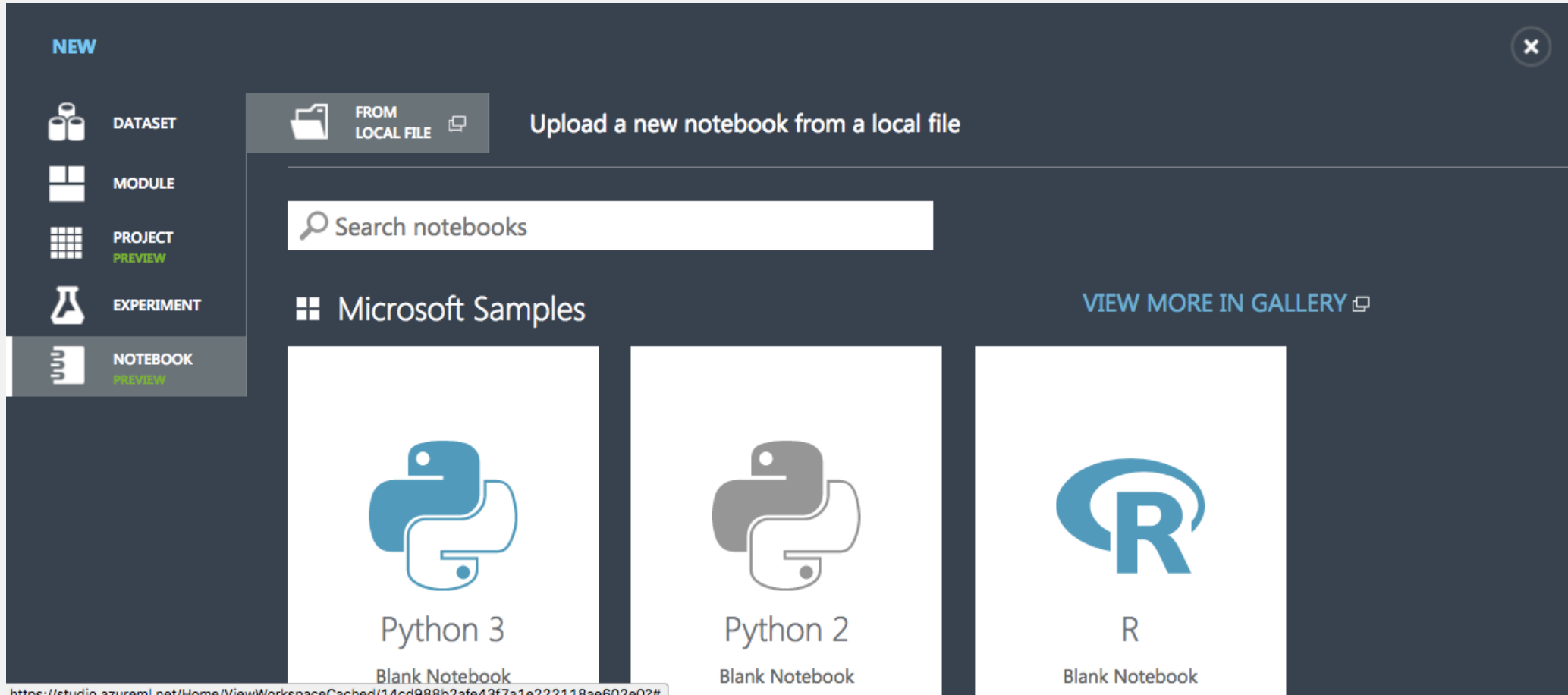


Microsoft



- Go to the following url
- <https://github.com/ahadmushir/MSDSBOOTCAMP/tree/master/Notebooks/module3>
- Download the notebooks.

Uploading an Ipython Notebook (Cont'd)



The screenshot displays the Microsoft Azure ML Studio interface. On the left, a sidebar titled 'NEW' contains icons and labels for 'DATASET', 'MODULE', 'PROJECT PREVIEW', 'EXPERIMENT', and 'NOTEBOOK PREVIEW'. The 'NOTEBOOK PREVIEW' option is highlighted. The main area shows a dialog titled 'Upload a new notebook from a local file' with a 'FROM LOCAL FILE' button. Below this, there is a search bar labeled 'Search notebooks'. A section titled 'Microsoft Samples' displays three notebook templates: 'Python 3', 'Python 2', and 'R', each with its respective logo and the text 'Blank Notebook'. A 'VIEW MORE IN GALLERY' link is visible on the right. The URL at the bottom of the browser window is <https://studio.azureml.net/Home/ViewWorkspaceCached/14cd988b2afe43f7a1e222118ae602e02#>.

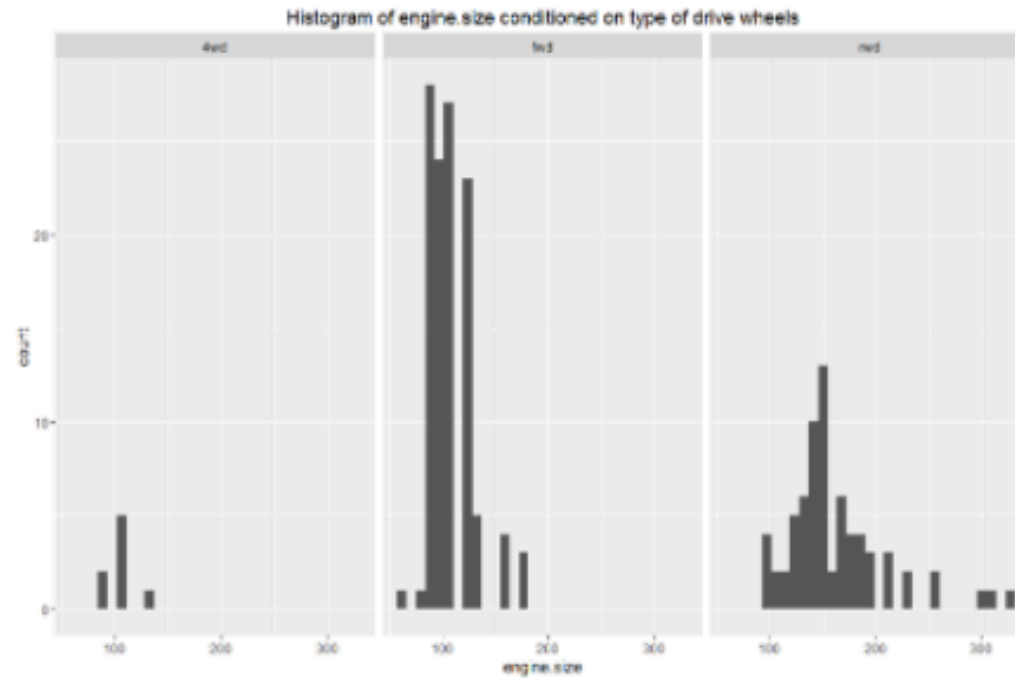


- Which of the following conditioned histograms most closely resembles the distribution of engine size conditioned on drive wheels?

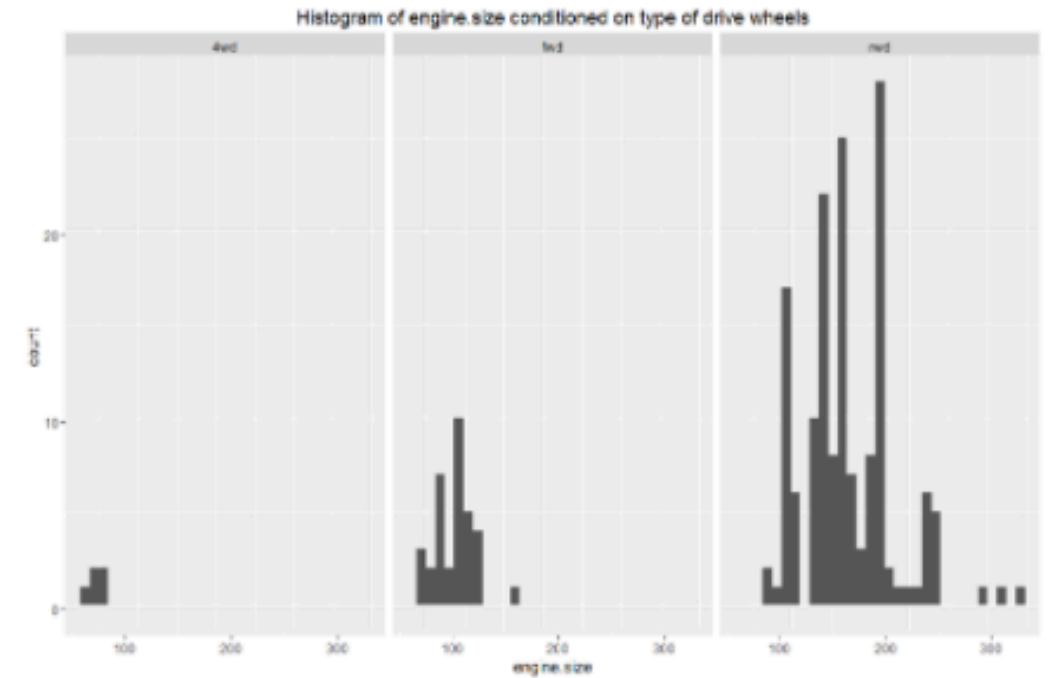
Questions from Hands On Lab



Microsoft



Histogram A

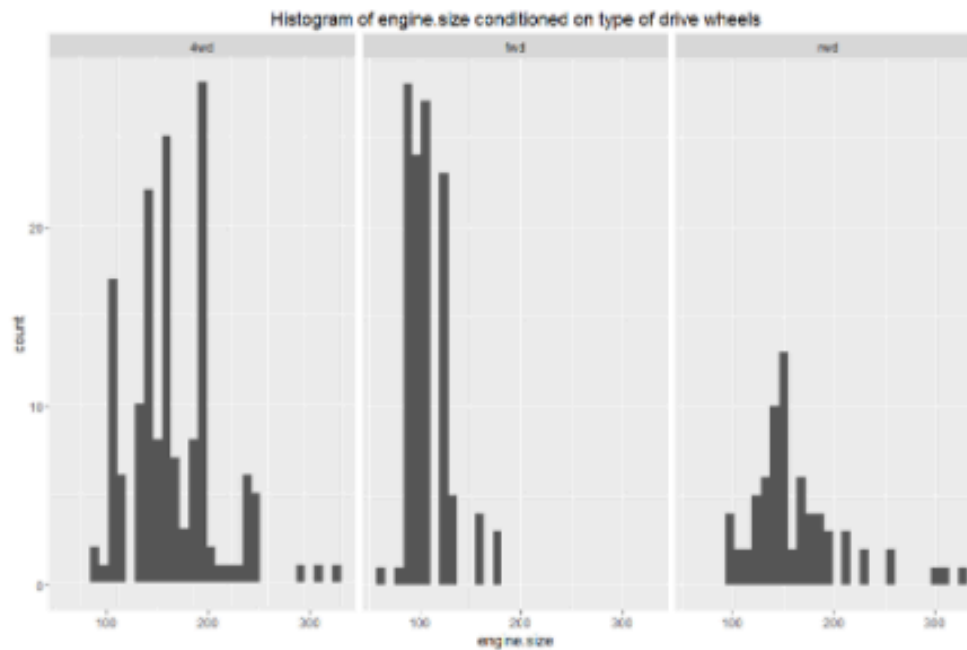


Histogram B

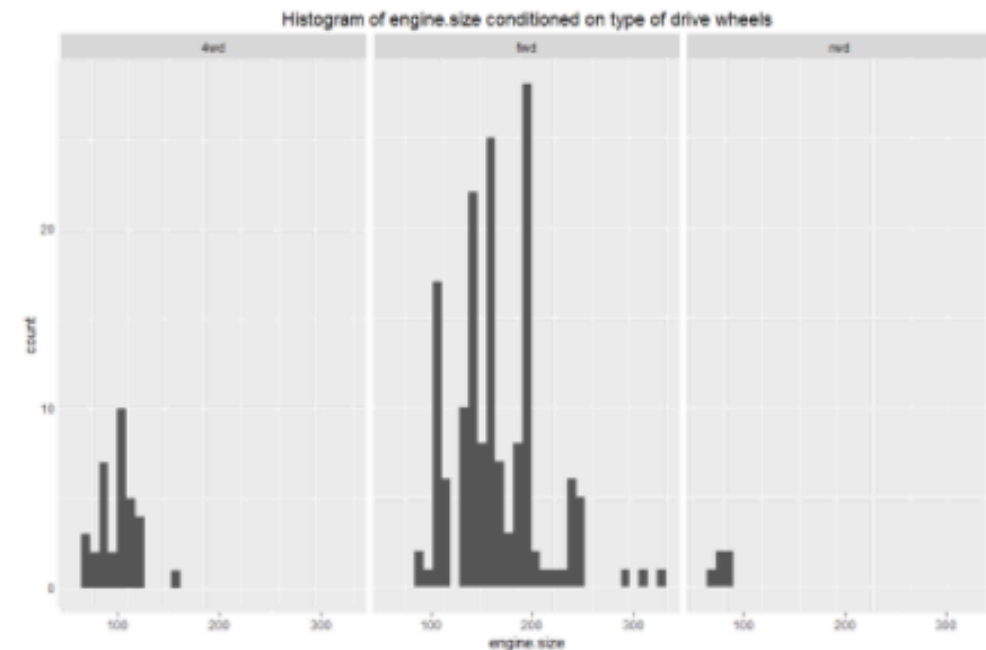
Questions from Hands On Lab



Microsoft



Histogram C



Histogram D



- **Plotting Adult Income**
- **Based on the conditioned box plots you created for the adult income classification dataset, the median age of adults who earn \$50K or less is...**
- A: Higher than the median age of adults who earn more than \$50K
- B: The same as the median age of adults who earn more than \$50K
- C: Lower than the median age of adults who earn more than \$50K

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Communications of ACM, 42:73-78, 1999.
- Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), December 1997.
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999.
- T. Redman. Data Quality: Management and Technology. Bantam Books, New York, 1992.
- Y. Wand and R. Wang. Anchoring data quality dimensions ontological foundations. Communications of ACM, 39:86-95, 1996.
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE Trans. Knowledge and Data Engineering, 7:623-640, 1995.

END OF MODULE 3
