PRACTICAL STATISTICS FOR DATA SCIENCE
REVIEW

EXPLORATORY DATA ANALYSIS

CHAPTER 1

Exploratory Data Analysis

EDA
EXPLORATORY DATA ANALYSIS

- First step in any data project: exploring the data.
- Classical statistics focused on inference:
  - complex set of procedures for drawing conclusions about large populations based on small samples
    - Testing hypothesis
    - Punctual and Interval estimation

ELEMENTS OF STRUCTURED DATA

- Data comes from many sources! Much of this data is unstructured.
- There are two basics types of structured data: **numeric and categorical.**

Numeric		
Continuous	Any value or interval	Interval, float, numeric
Discrete	Only integer values	Integer, count
Categorical		
Categorical	Categories	Enums, factors, nominal
Binary	Two categories	Dichotomous, logical, Boolean
Ordinal	Categorical that has explicit order	Ordered factor

ELEMENTS OF STRUCTURED DATA

Knowing the data type is important to help determine that type of visual display, data analysis or statistical model.

Data science software, (R/python) uses these data types to improve computational performance. The data type for a variable determines how software will handle computations for that variable.

RECTANGULAR DATA

- The typical frame of reference for an analysis in data science is a rectangular data object, like a spreadsheet or database table
- Data frame:
  - Rectangular data is the basic data structure for statistical and machine learning models
- Feature:
  - A column in the table is commonly referred to as a feature. (attribute, input, predictor, independent variable)
- Outcome:
  - The features are sometimes used to predict the outcome. (dependent variable, response, target, output)
- Records:
  - A row in the table is commonly referred to as a record. (case, example, instance, observation, pattern, sample)

DATA FRAMES AND INDEXES

R

- data.frame object
- An automatic index is created for a data.frame based on the order of the rows.
- Doesn't support multilevel indexes. To overcome this use data.table or dplyr libraries.

PYTHON

- DataFrame object (pandas).
- An automatic index is created for a DataFrame based on the order of the rows.
- Pandas can handle multiple indexes.

NONRECTANGULAR DATA STRUCTURES

- Time series records successive measurements of the same variable.
- Spatial data structures, used in mapping and location analytics.
- Graph (network) used to represent physical, social and abstract relationships.

ESTIMATES

- **Estimates:** values calculated from the data at hand, to draw a distinction between what we see from the data and the theoretical or true value. Data scientist and business analyst are more likely to refer to such values as **metric**.
- Bias: difference between the expected value of an estimator and the true value.

$$Bias_{\theta} = E\left[\hat{\theta}\right] - \theta$$

- If the bias of an estimator is 0, we have an unbiassed estimator.

ESTIMATES OF LOCATION

- A basic step in exploring data is getting a "**typical value**" for each feature an estimate of where most of the data is located.
- **Mean** (average): sum of all values divided by the number of value.

$$\bar{x} = \frac{\sum_i^n x_i}{n}$$

*probe that the mean is an unbiassed estimator.

- Weighted mean: sum of all values times a weight divided by the sum of weights.
  - When some values are intrinsically more variable than others, and highly variable observations are given a lower weight.

$$\bar{x}_w = \frac{\sum_i^n w_i x_i}{\sum_i^n w_i}$$

- Trimmed mean: the average of all values after dropping a fixed number of extreme values.
  - Eliminates the influence of extreme values.

$$\bar{x} = \frac{\sum_{i=p}^{n-p} x_{(i)}}{n - 2p}$$

ROBUST ESTIMATES

- **Robust**: Not sensitive to extreme values
- **Outlier**: A data value that is very different from most of the data
- **Median (50th percentile)**: The value such that one half of the data lies above and below.
- **Weighted median**: The value such that one half of the sum of the weights lies above and below the sorted data.

ESTIMATES OF VARIABILITY

- **Variability**, also referred to as dispersion, measures whether the data values are tightly clustered or spread out.
- **Deviations**: The difference between the observed values and the estimate of location (errors, residuals)

$$e_i = x_i - \bar{x}$$

- **Variance**: the sum of squared deviations from the mean divided by n-1 where n is the number of data values.

$$s^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$$

- ***probe that the variance is an unbiassed estimator.**

- **Standard deviation:** The square root of the variance
    - Is much easier to interpret than the variance since it is in on the same scale as the original data.
- **Mean absolute deviation:** The mean of the absolute value of the deviation from the mean.

$$mean\ absolute\ deviation = \frac{\sum_i^n |x_i - \bar{x}\,|}{n}$$

- **Median absolute deviation (MAD):** The median of the absolute value of the deviation from the median.

$$MAD = Median(|x_1 - m|, |x_2 - m|, \dots, |x_N - m|)$$

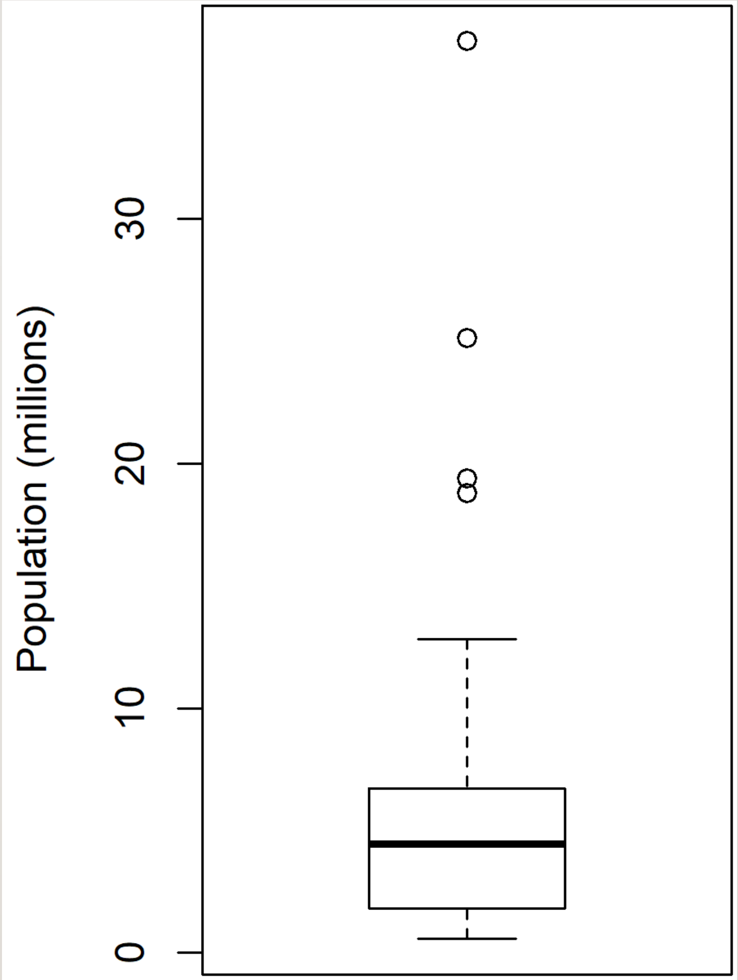
ESTIMATED BASED ON PERCENTILES

- A different approach to estimating dispersion is based on looking at the spread of the sorted data. Also known as **order statistics**.
- **Range**: The difference between the largest and the smallest value in a data set. (ranks)
- **Percentile**: The value such that P percent of the values take on this value or less and (100-P) percent take on this value or more. (quantile)
- Interquartile range: The difference between the 75th percentile and the 25th percentile. (IQR)

EXPLORING THE DATA DISTRIBUTION

- Estimators are useful to explore how data is distributed, overall.
- There are a groups of tools that provide us insights about the data distribution:
    - Boxplot (box and whiskers plot)
    - Frequency table
    - Histogram
    - Density plot

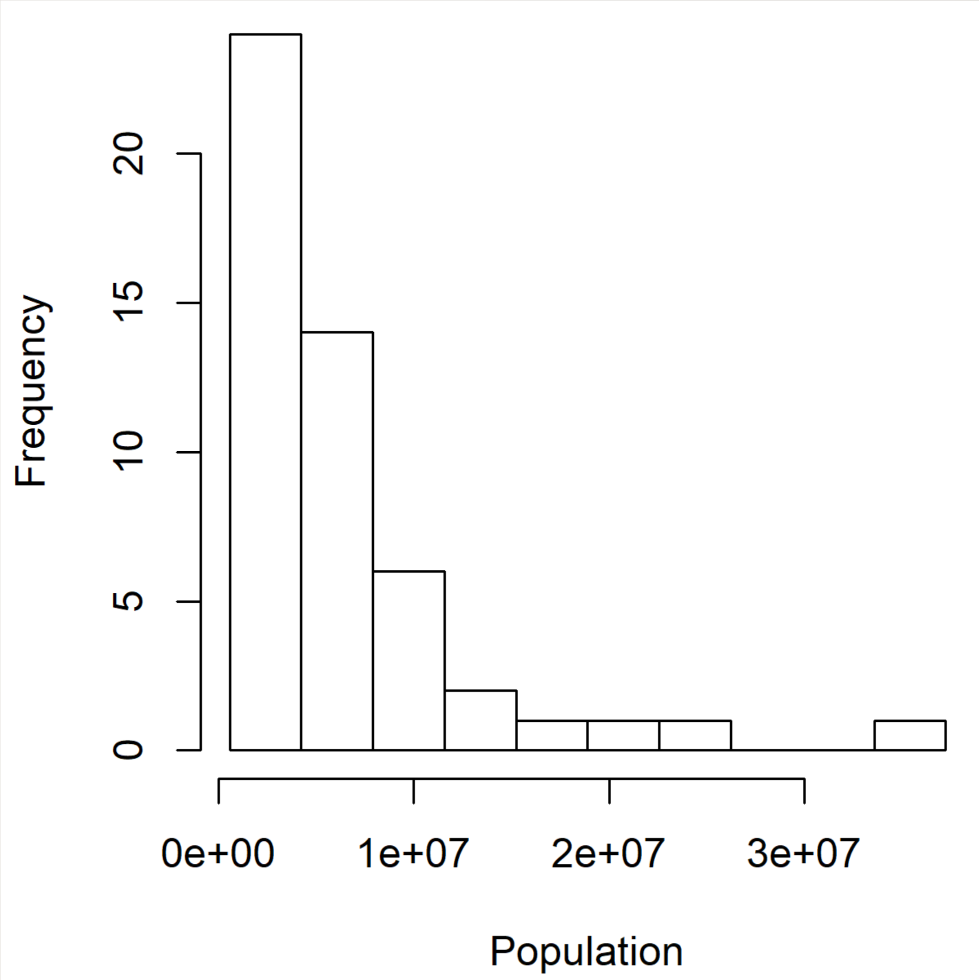
PERCENTILES AND BOXPLOTS



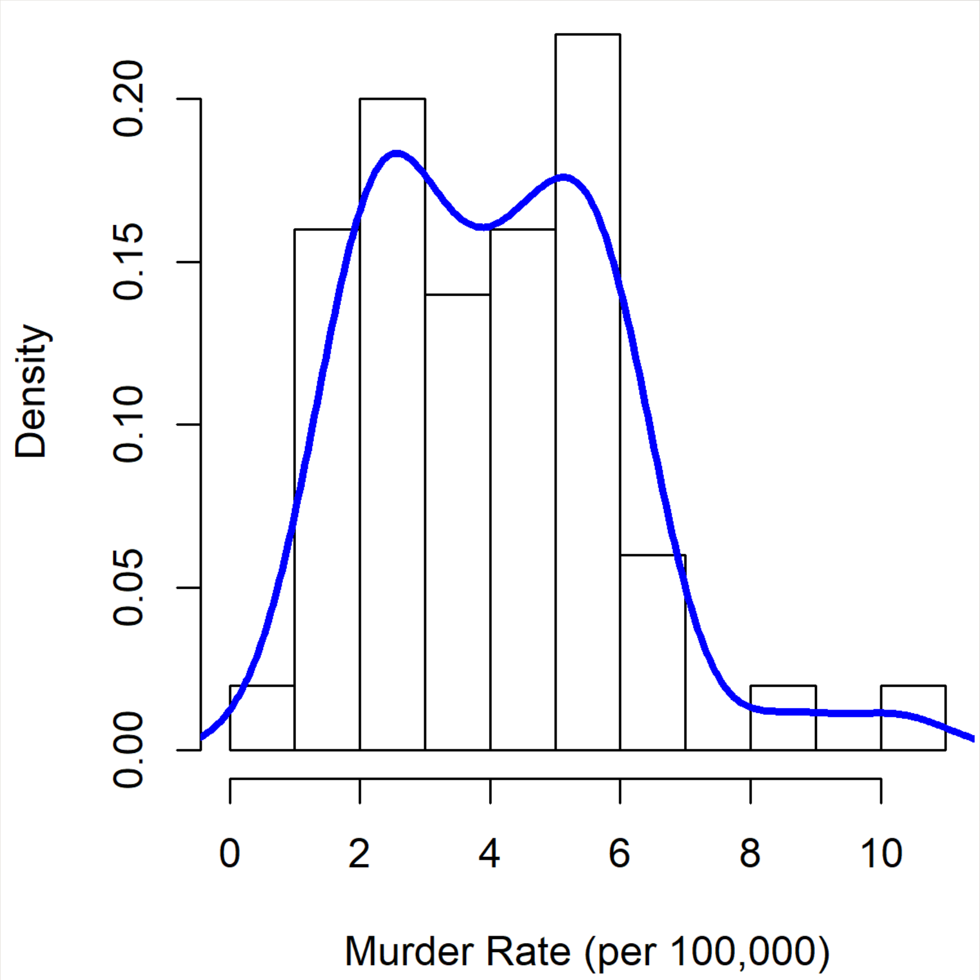
FREQUENCY TABLES

BinNumber	BinRange	Count	States
1	563,626-4,232,658	24	WY,VT,ND,AK,SD,DE,MT,RI,NH,ME,HI,ID,NE,WV,NM,NV,UT,K...
2	4,232,659-7,901,691	14	KY,LA,SC,AL,CO,MN,WI,MD,MO,TN,AZ,IN,MA,WA
3	7,901,692-11,570,724	6	VA,NJ,NC,GA,MI,OH
4	11,570,725-15,239,757	2	PA,IL
5	15,239,758-18,908,790	1	FL
6	18,908,791-22,577,823	1	NY
7	22,577,824-26,246,856	1	TX
8	26,246,857-29,915,889	0	
9	29,915,890-33,584,922	0	
10	33,584,923-37,253,956	1	CA

HISTOGRAMS



DENSITY ESTIMATES

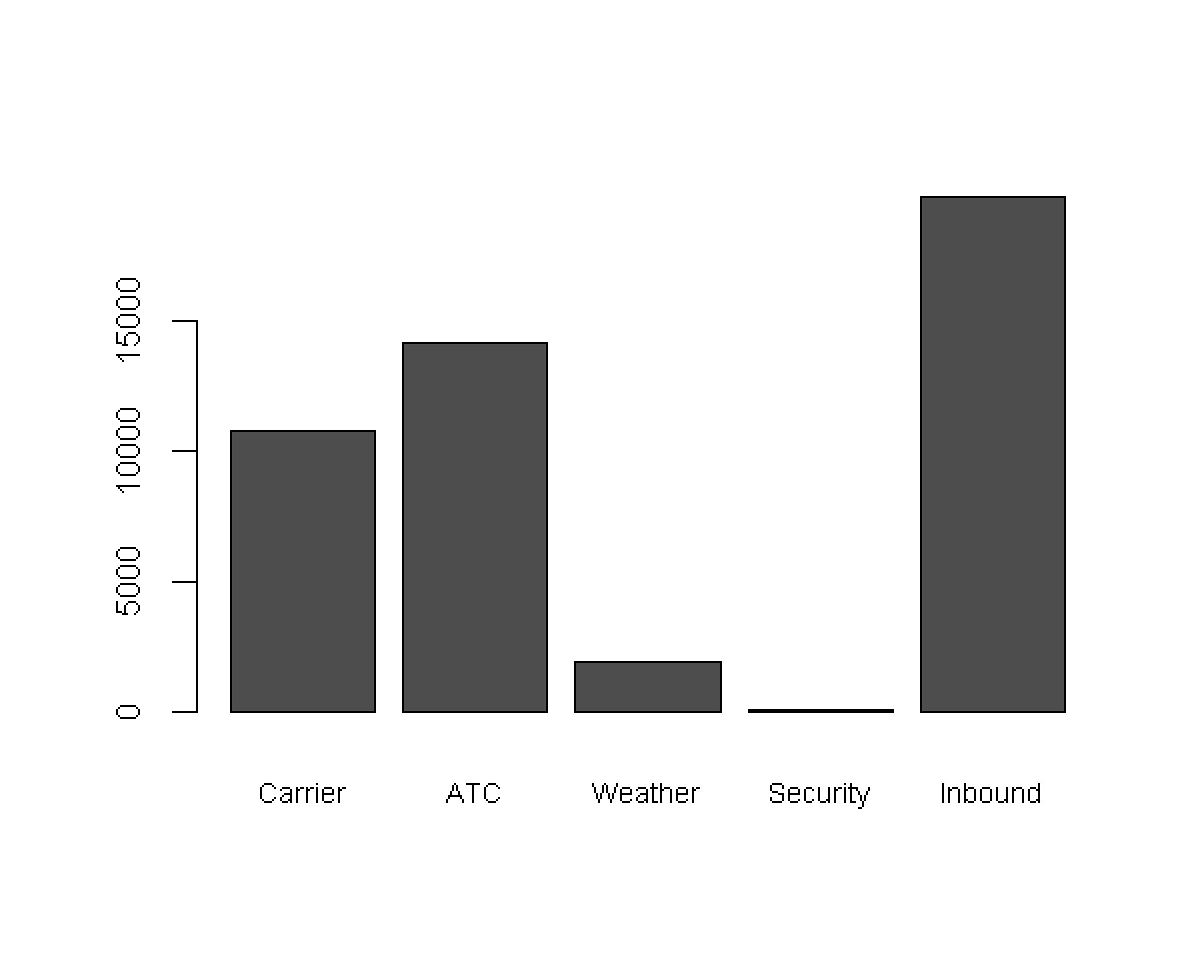


EXPLORING BINARY AND CATEGORICAL DATA

- For categorical data, simple proportions or percentages tell the story of the data.
- **Mode**: The most commonly occurring category or value in a data set.
- **Expected value**: When the categories can be associated with a numeric value, this gives an average value based on a category's probability of occurrence.
- **Bar charts:** The frequency or proportion for each category plotted as bars.
- **Pie charts**: The frequency or proportion for each category plotted as wedges in a pie.

BAR CHART

Bar chart resembles a histogram' in a bar chart the x-axis represents different categories of a factor variable, while in a histogram the x-axis represents values of a single variable on a numeric scale.



EXPECTED VALUE

- Expected value is the sum of values times their probability of occurrence, often used to sum up factor variable levels.
- Example: A new cloud technology offers two levels of service. Service A is priced at $300/month and service B at $50/month. 5% of webinar attendees will sign up for the $300 service, 15% for the $50 service and %80 will not sign up for anything.

$$EV = (0.05)(300) + (0.15)(50) + (0.80)(0) = 22.5$$

CORRELATION

- Correlation coefficient: A metric that measures the extent to which numeric variables are associated with one another (ranges from -1 to 1)

$$r = \frac{\sum_1^N (x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

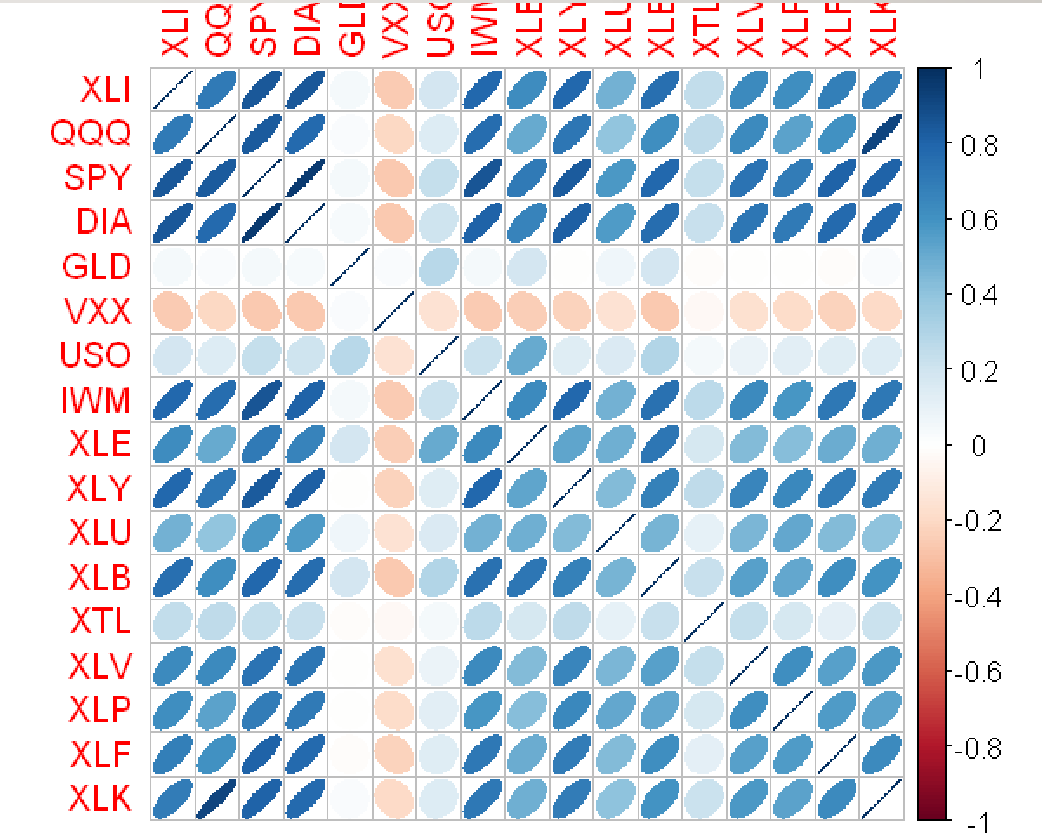
*Variables can have an association that is not linear, in which case the correlation coefficient may not be a useful metric.

** Correlation coefficient is sensitive to outliers in the data.

*** The definition above corresponds to Pearson's correlation definition.

- Correlation matrix: A table where the variables are shown on both rows and columns, and the cell values are the correlations between the variables.

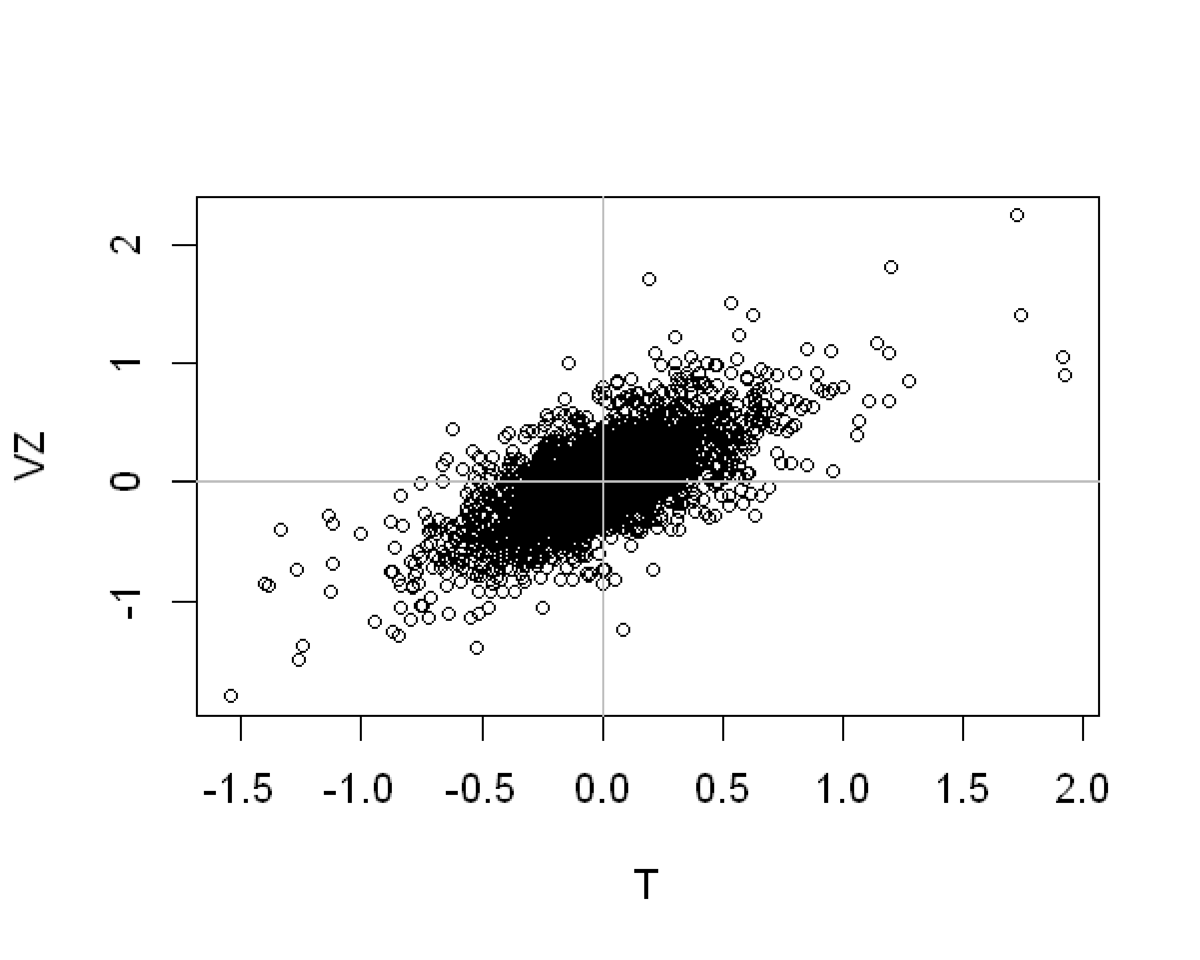
CORRPLOT



XLI
QQQ
SPY
DIA
GLD
VXX
USO
IWM
XLE
XLY
XLU
XLB
XTL
XLV
XLP
XLF
XLK
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

SCATTERPLOTS

The standard way to visualize the relationship between two measured data variables is with a scatterplot

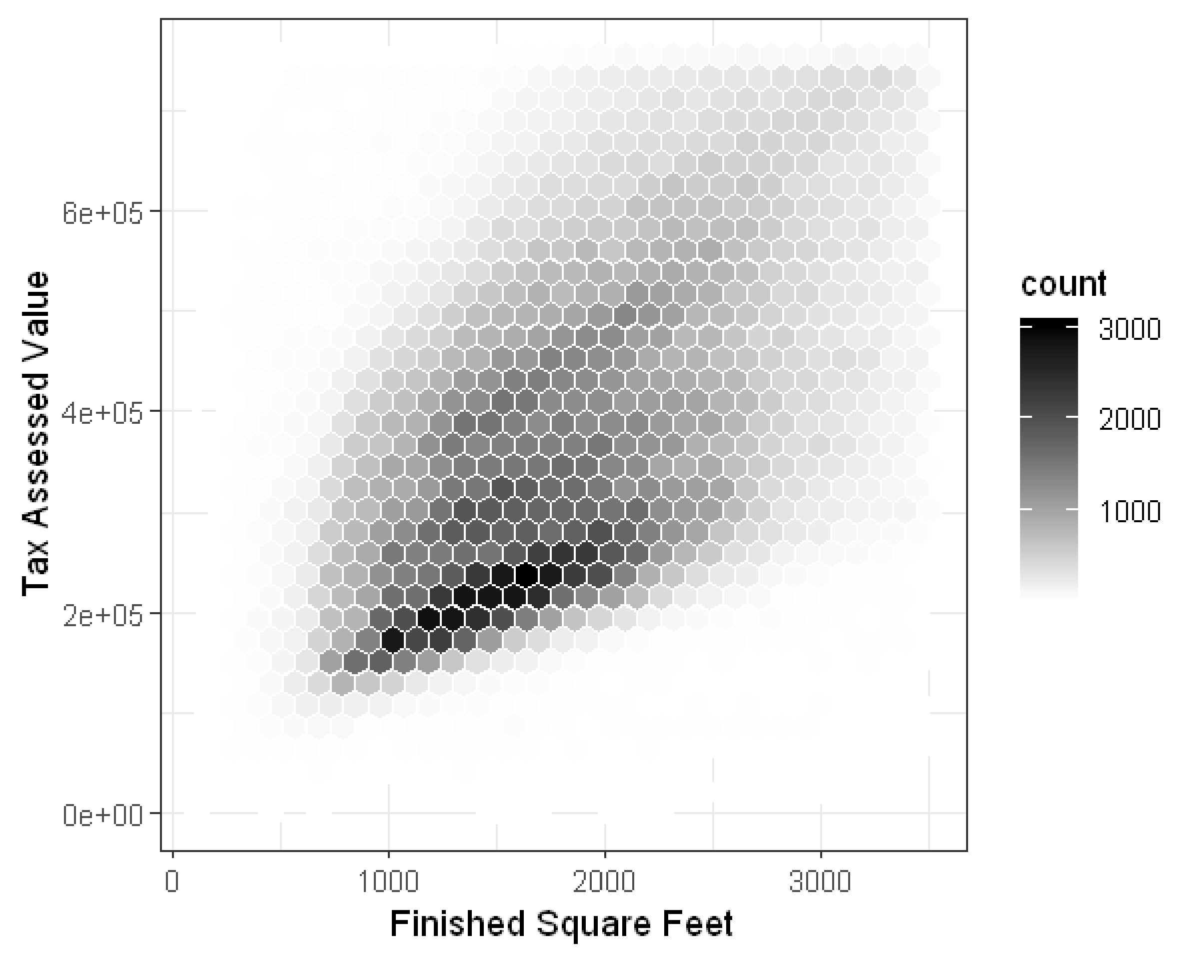


EXPLORING TWO OR MORE VARIABLES

- Multivariate exploratory analysis tools:
- **Hexagonal binning**: A plot of two numeric variables with the records binned into hexagons.
- **Contour plots**: A plot showing the density of two numeric variables like a topographical map.
- **Violin plots**: Similar to a boxplot but showing the density estimate.

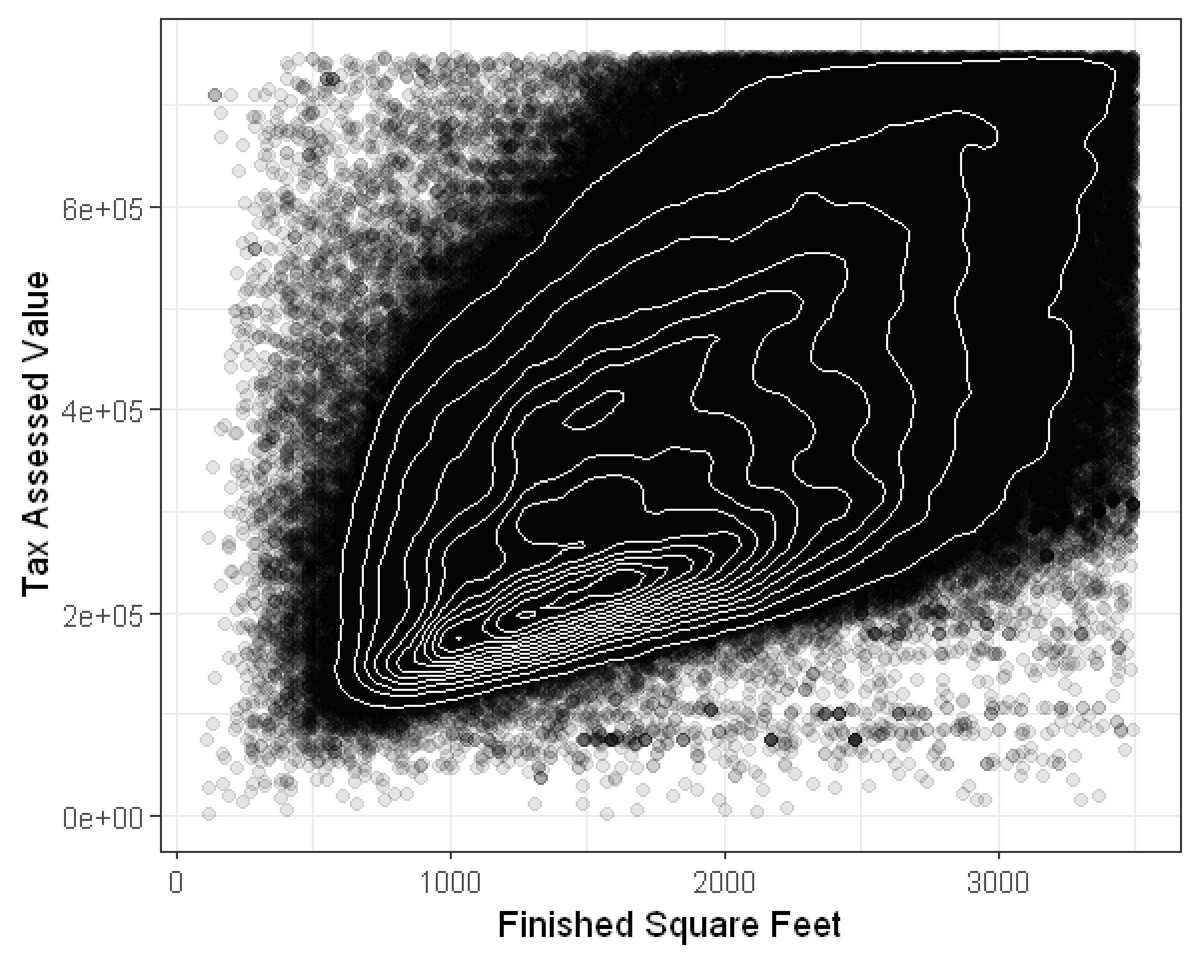
HEXAGONAL BINNING

For dataset with hundreds of thousands or millions od records, a scatterplot will be too dense, so we need a different way to visualize the relationship. Dots are grouped into hexagonal bins a plotted the hexagons with a color indicating the number of records in that bin.



CONTOURS AND HEATMAPS

We can overlay a contour on a scatterplot to visualize the relationship between two numeric variables. The contours are essentially a topographical map to two variables; each contour band represent a specific density of points, increasing as one nears a "peak".



TWO CATEGORICAL VARIABLES

Contingency tables: A tally of counts between two or more categorical variables.

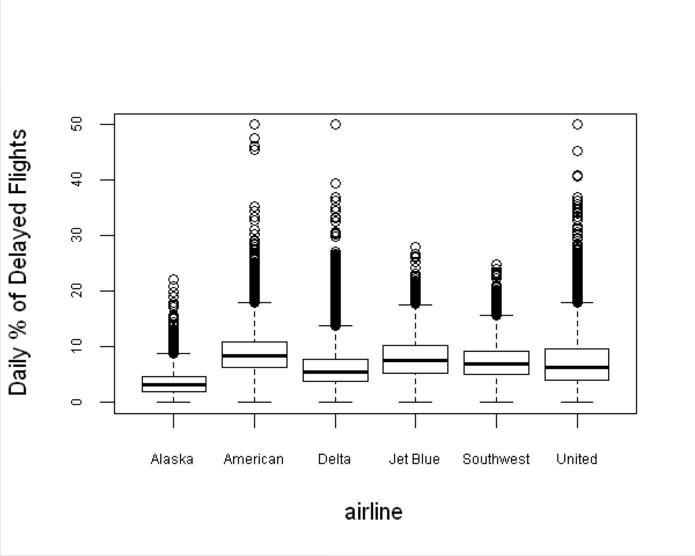
Grade	Charged Off	Current	Fully Paid	Late	Total
A	1562	50051	20408	469	72490
	0.022	0.690	0.282	0.006	0.161
B	5302	93852	31160	2056	132370
	0.040	0.709	0.235	0.016	0.294
C	6023	88928	23147	2777	120875
	0.050	0.736	0.191	0.023	0.268
D	5007	53281	13681	2308	74277
	0.067	0.717	0.184	0.031	0.165
E	2842	24639	5949	1374	34804
	0.082	0.708	0.171	0.039	0.077
F	1526	8444	2328	606	12904
	0.118	0.654	0.180	0.047	0.029
G	409	1990	643	199	3241
	0.126	0.614	0.198	0.061	0.007
Total	22671	321185	97316	9789	450961

CATEGORICAL AND NUMERIC DATA

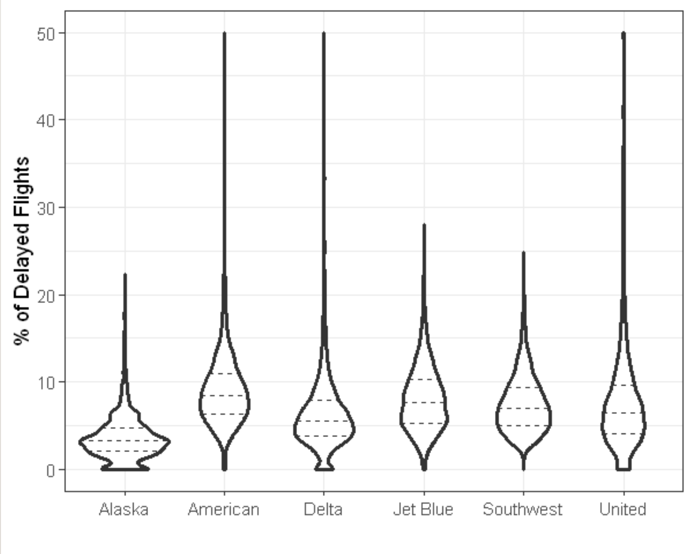
- Boxplots are a simple way to visually compare the distribution of a numeric variable grouped according to a categorical variable.
- A violin plot is an enhancement to the boxplot and plots the density estimate with the density on the y-axis. The advantage of a violin plot is that it can show nuances in the distribution that aren't receptible in a boxplot.

CATEGORICAL AND NUMERIC DATA

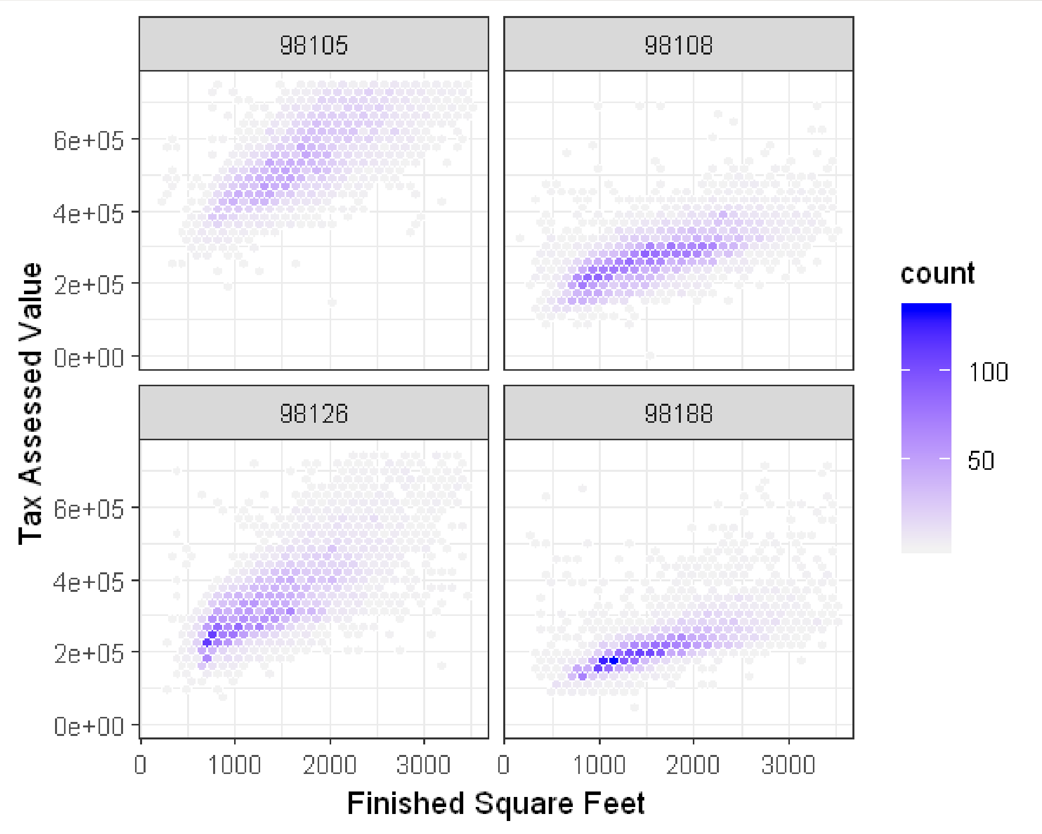
BOXPLOT



VIOLIN PLOT



VISUALIZING MULTIPLE VARIABLES



SUMMARY

- Key idea of EDA: Look at the Data!!!
- By summarizing and visualization the data, you can give valuable intuition and understanding of the project.
- EDA should be a cornerstone of any data science project