Exploratory Data Analysis

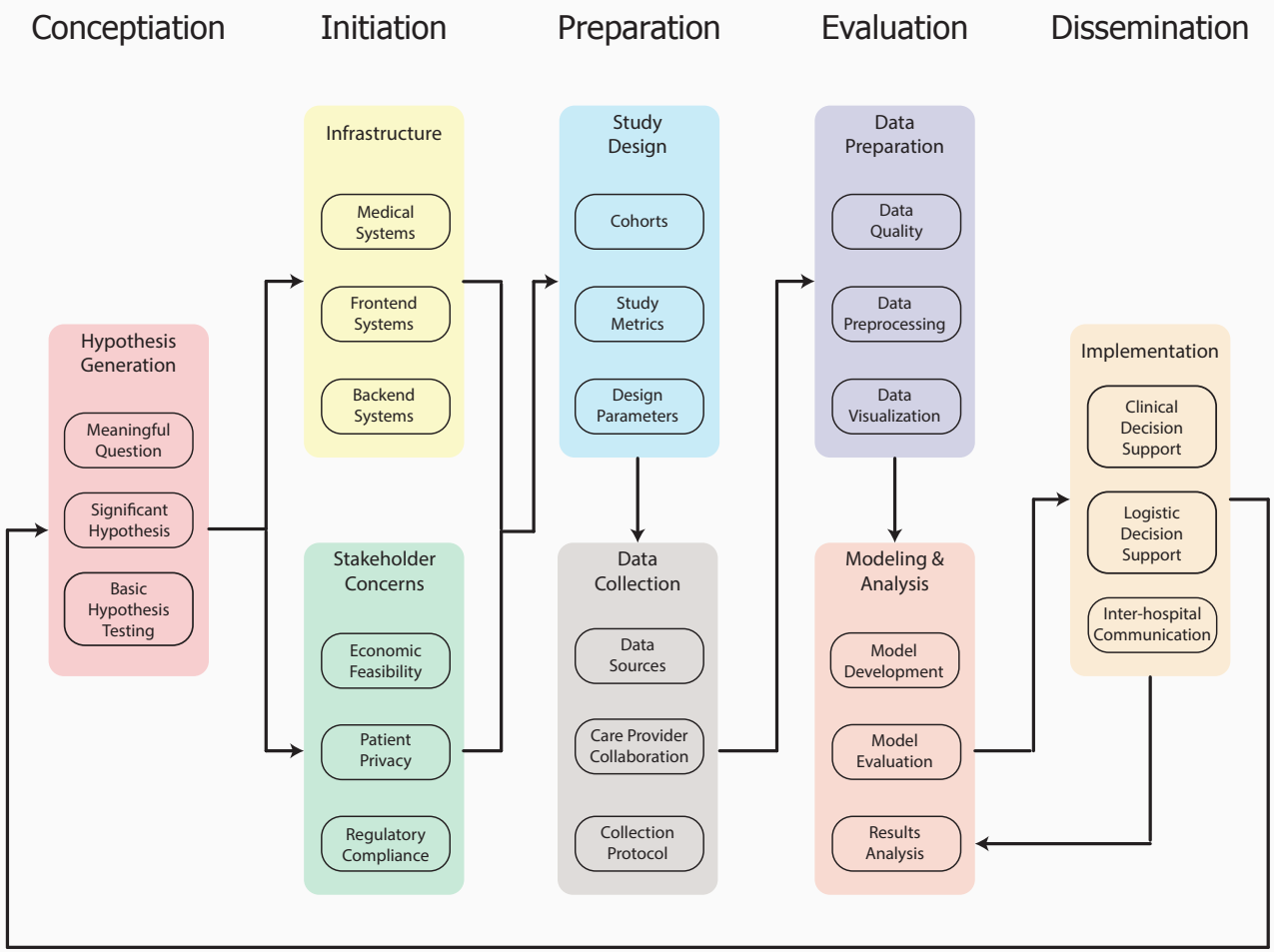
Digital Transformation of Healthcare

Michoel Snow M.D. Ph.D. and Glen Ferguson Ph.D.

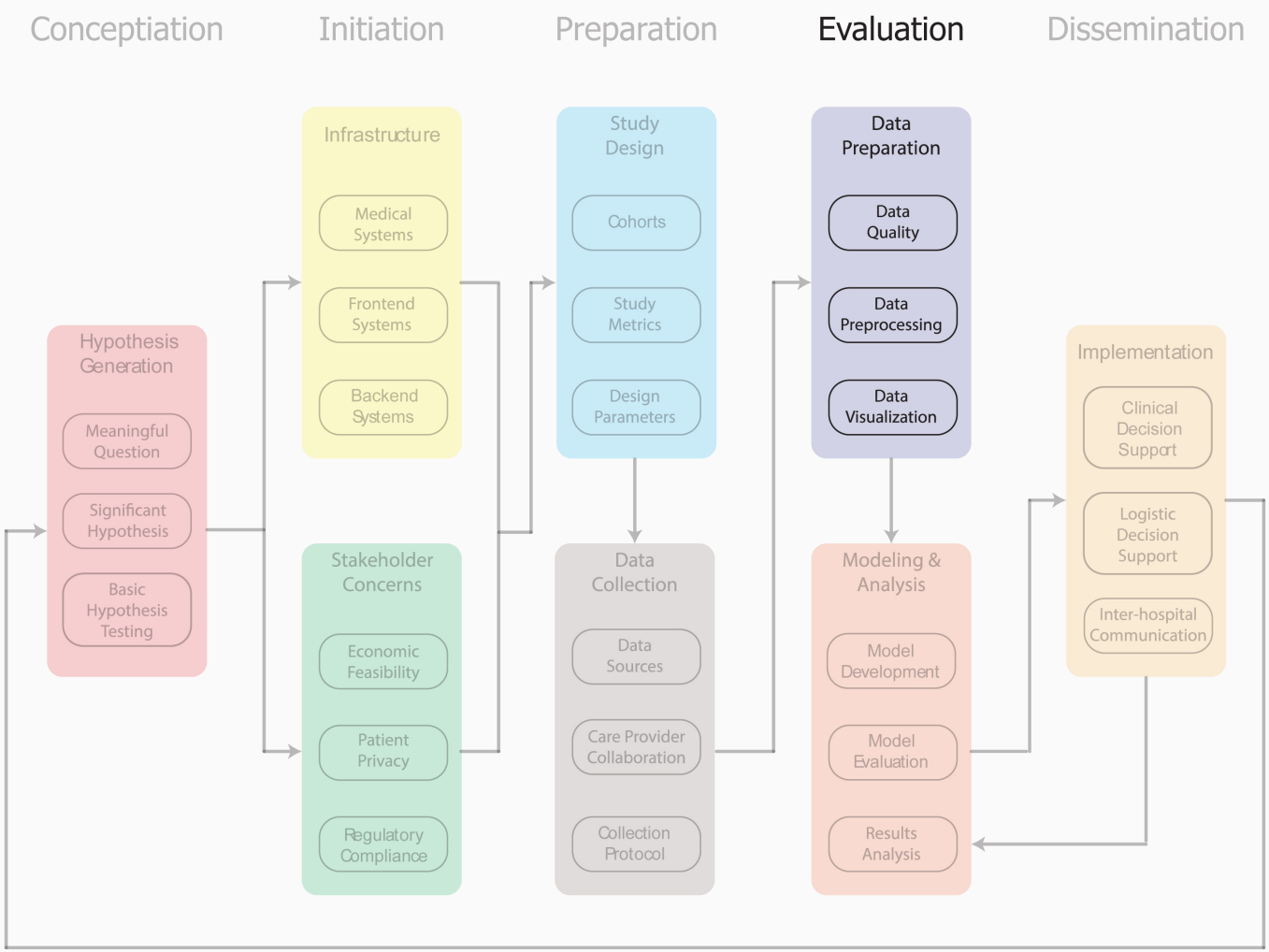
Center for Health Data Innovations

Exploratory Data Analysis

Bioinformatics Pipeline



Data Analysis



EDA

- Objectives
  - Define EDA
  - Know the purpose of EDA
  - Understand a visualization toolbox
  - Ask questions about data using EDA

What is EDA?

- EDA

What is EDA?

- EDA
  - An method of summarizing the main points of a data set, most often using visualization

What is EDA?

- EDA
  - An method of summarizing the main points of a data set, most often using visualization
  - Distinct from other types of analysis for confirming or validating hypotheses

What is EDA?

- EDA
  - An method of summarizing the main points of a data set, most often using visualization
  - Distinct from other types of analysis for confirming or validating hypotheses
  - Used to find hidden structure in the data, e.g., unknown relations between the variables or correlation with the target variable

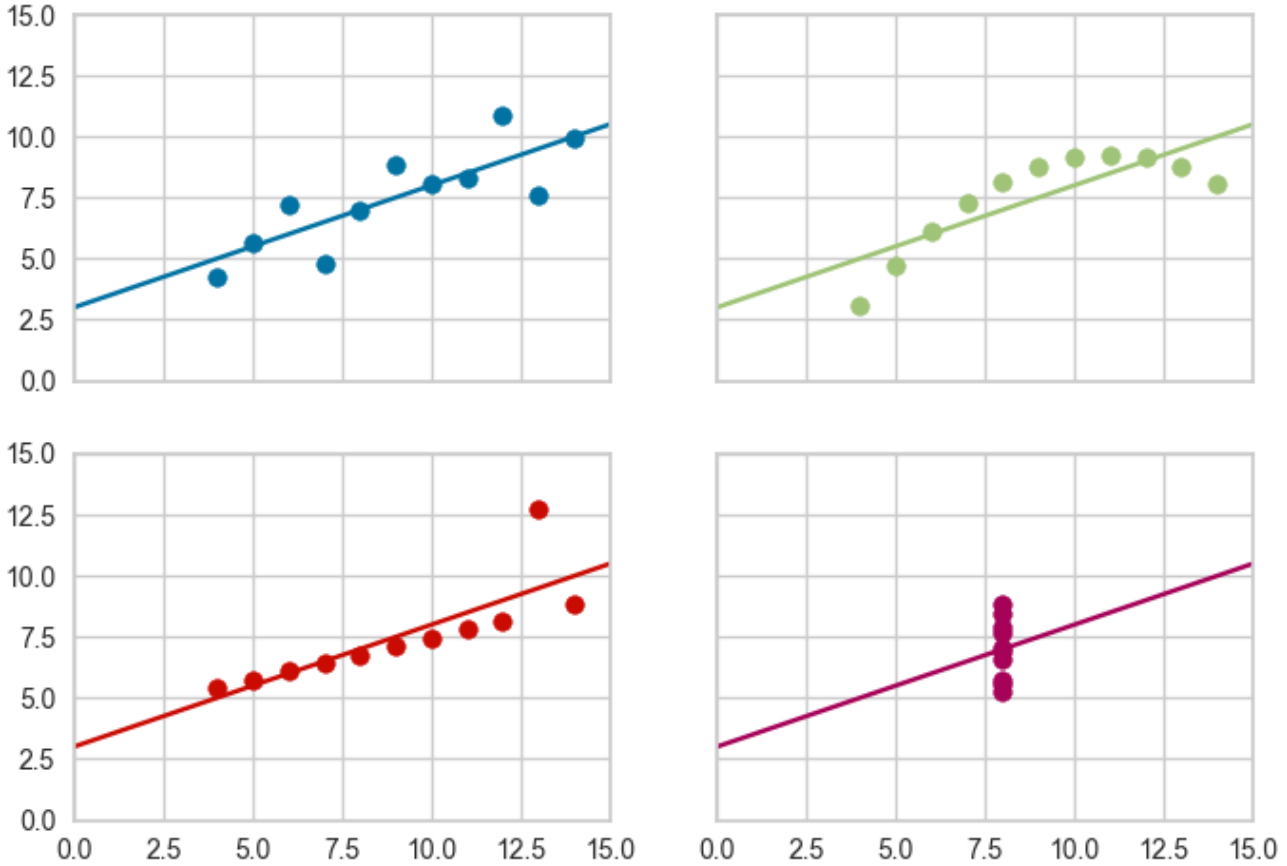
Why visualize?

Why not just use summary statistics?

Why visualize?

Why not just use summary statistics?

Figure 1: Anscombe's Quartet



Why perform EDA?

- EDA

Why perform EDA?

- EDA
  - Assess data quality

Why perform EDA?

- EDA
  - Assess data quality
  - Understand the types of data in the data set
    - Continuous
    - Time Series
    - Categorical
    - Boolean (T/F)
    - Low Cardinality (few categories)
    - High Cardinality (few categories)

Why perform EDA?

- EDA
  - Assess data quality
  - Understand the types of data in the data set
    - Continuous
    - Time Series
    - Categorical
    - Boolean (T/F)
    - Low Cardinality (few categories)
    - High Cardinality (few categories)
  - Determine the parameters of the data set
    - Min, max, median, and percentiles of numerical data
    - Distribution of the numerical data
    - Count, number of unique values, most common value, and frequency of most common values for categorical values
    - Range and rolling values of a time series

Why perform EDA?

- EDA
  - Assess data quality
  - Understand the types of data in the data set
    - Continuous
    - Time Series
    - Categorical
    - Boolean (T/F)
    - Low Cardinality (few categories)
    - High Cardinality (few categories)
  - Determine the parameters of the data set
    - Min, max, median, and percentiles of numerical data
    - Distribution of the numerical data
    - Count, number of unique values, most common value, and frequency of most common values for categorical values
    - Range and rolling values of a time series
  - Develop hypothesis

Why perform EDA?

- EDA
  - Assess data quality
  - Understand the types of data in the data set
    - Continuous
    - Time Series
    - Categorical
    - Boolean (T/F)
    - Low Cardinality (few categories)
    - High Cardinality (few categories)
  - Determine the parameters of the data set
    - Min, max, median, and percentiles of numerical data
    - Distribution of the numerical data
    - Count, number of unique values, most common value, and frequency of most common values for categorical values
    - Range and rolling values of a time series
  - Develop hypothesis
  - Determine what prepossessing and modeling are appropriate

Summary Stats

Summary Stats on Appt. Attendance Data Set

	ApptTime	LeadDays	PatientAge
mean	11.663151	44.736143	44.722064
std	2.299348	35.997952	22.396794
min	8.100000	-3.000000	0.000000
25%	9.500000	15.000000	26.000000
50%	11.100000	35.000000	46.000000
75%	14.000000	70.000000	62.000000
max	16.300000	369.000000	104.000000

Summary Stats

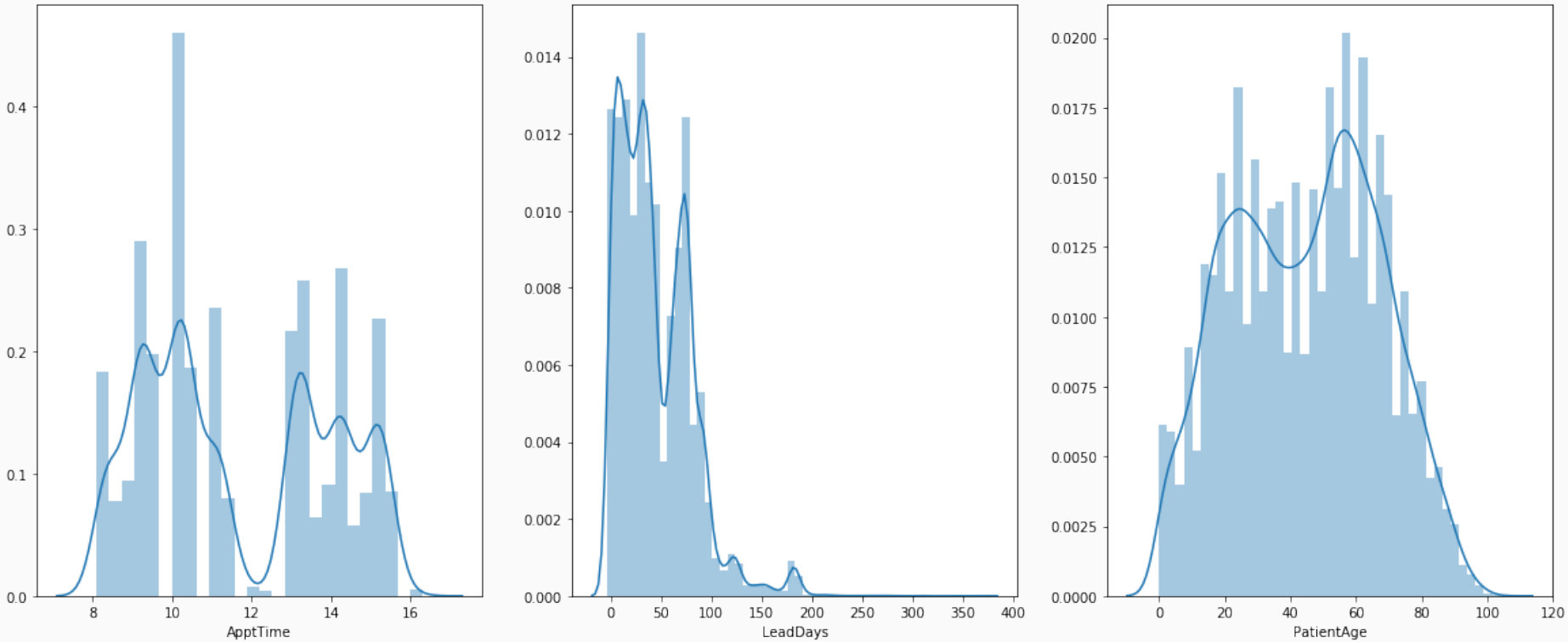
Summary Stats on Appt. Attendance Data Set

	ApptMonth	ApptDays	AppointmentBlock
count	21451	21451	21451
unique	12	5	5
top	May	Tue	None
freq	2023	4971	20932

Visualization Toolbox

Univariate exploration

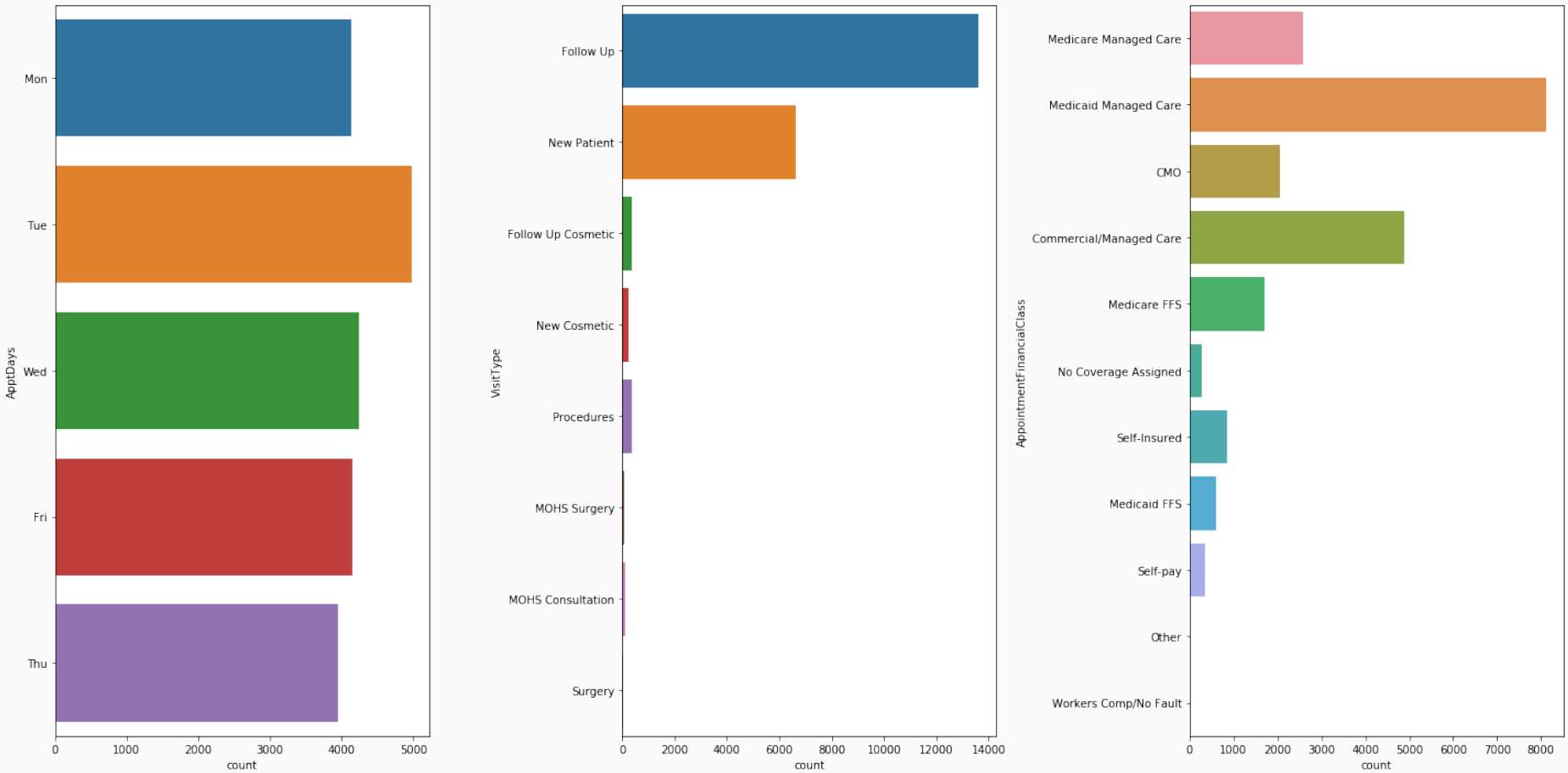
Figure 2: Distributions of continuous variables



Visualization Toolbox

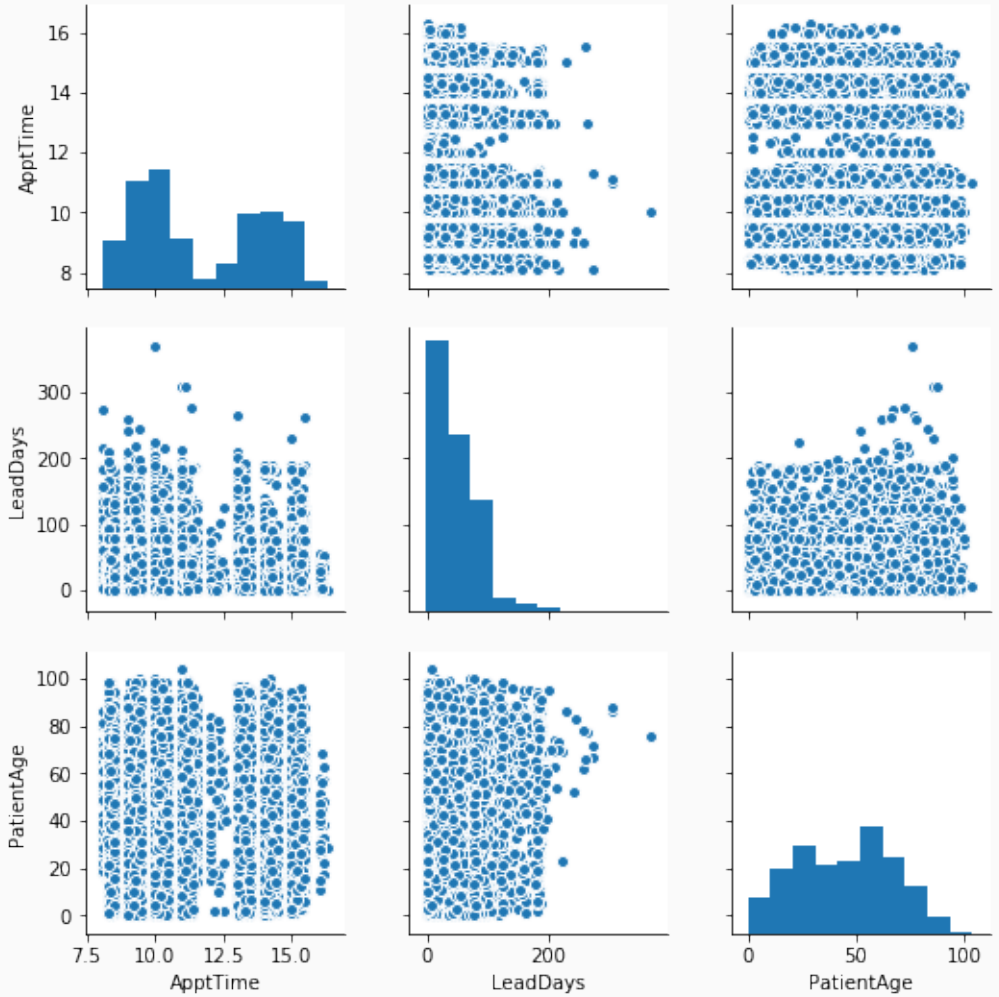
Univariate exploration

Figure 3: Counts of discreet variables



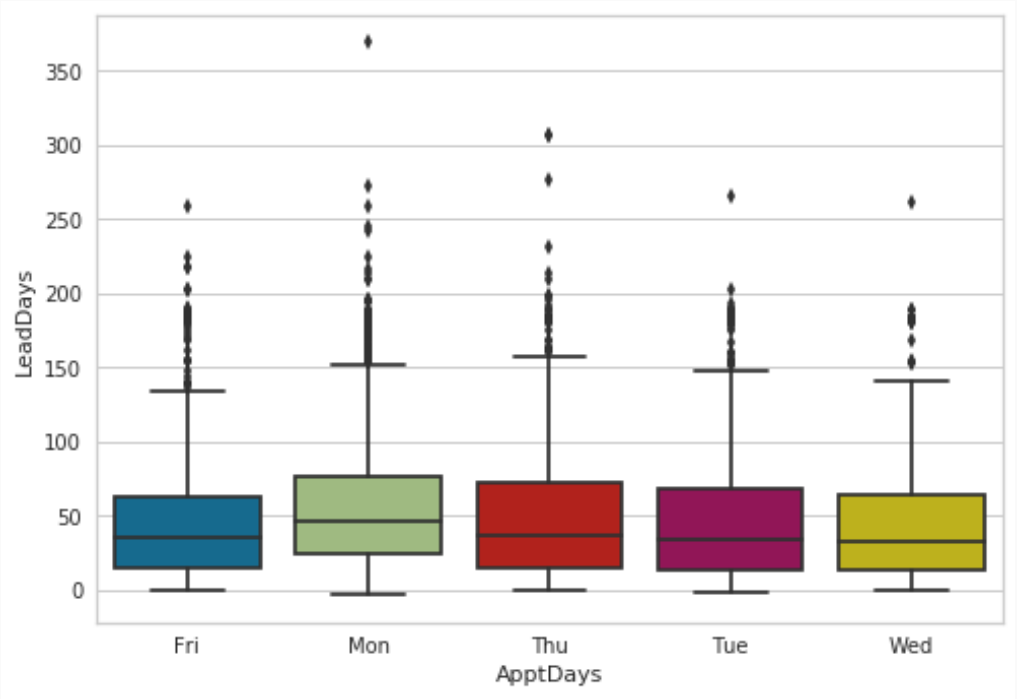
Bivariate exploration

Figure 4: Pair plot of continuous variables



Bivariate exploration

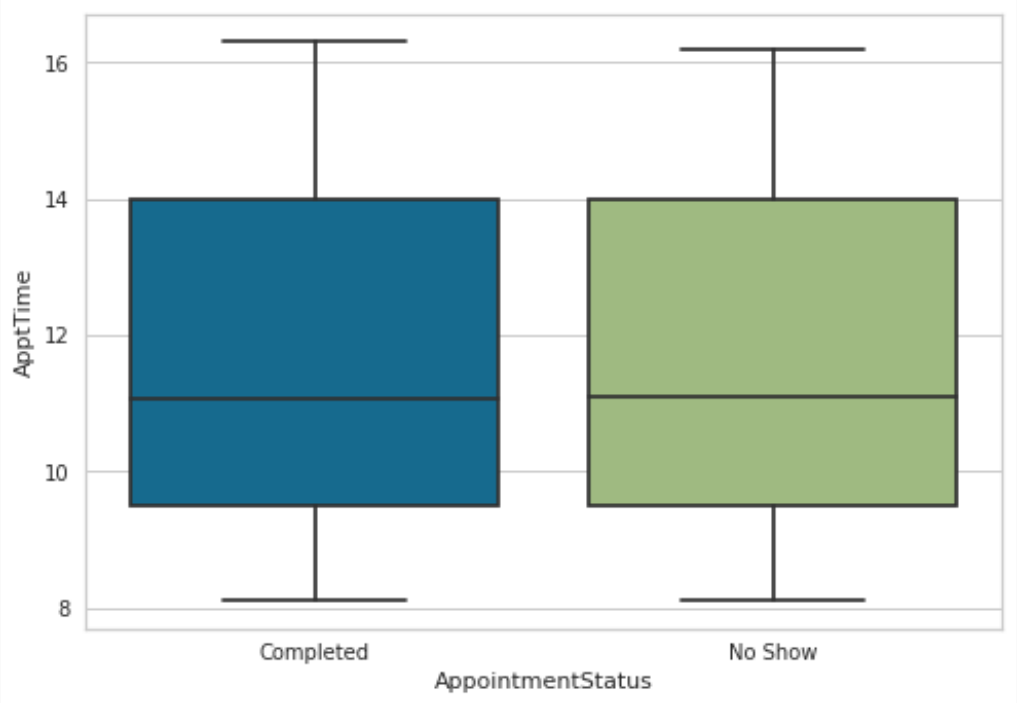
Figure 5: Box plots discrete-continuous variables



Visualization Toolbox

Bivariate exploration

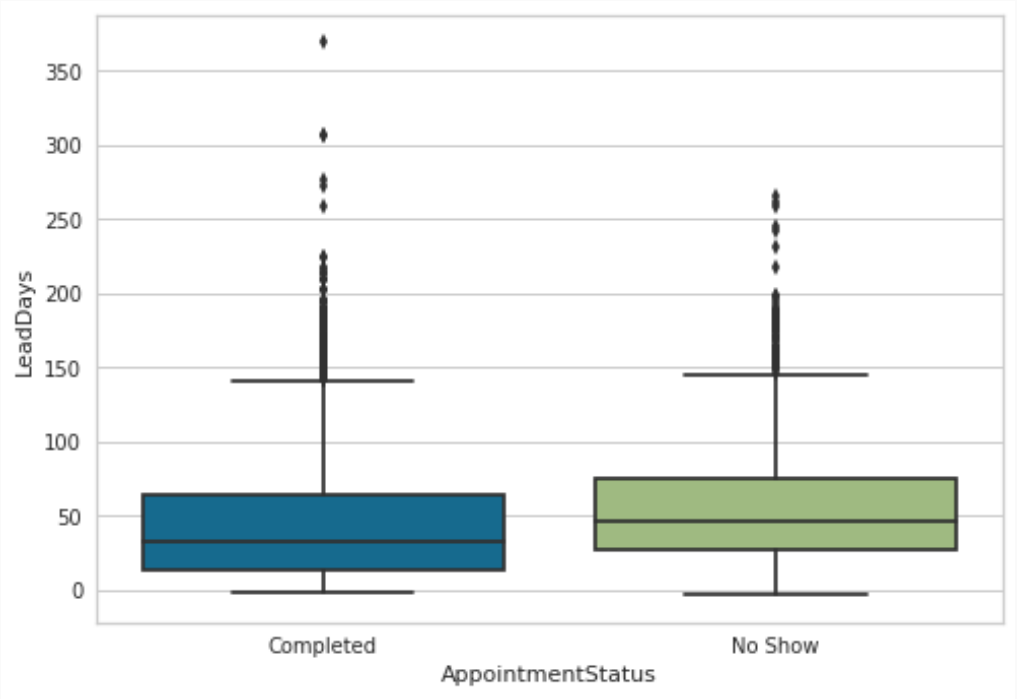
Figure 6: Box plots discrete-continuous variables



Visualization Toolbox

Bivariate exploration

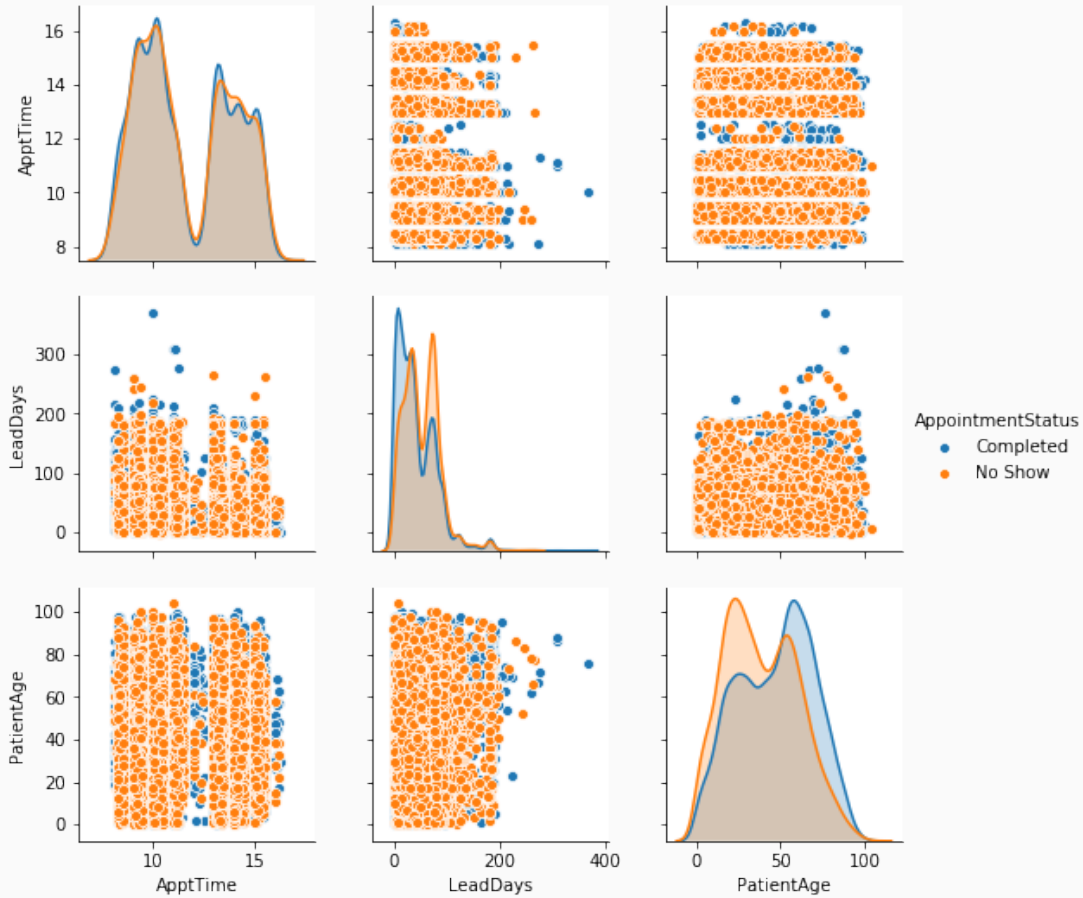
Figure 7: Box plots discrete-continuous variables



Visualization Toolbox

Bivariate exploration with target

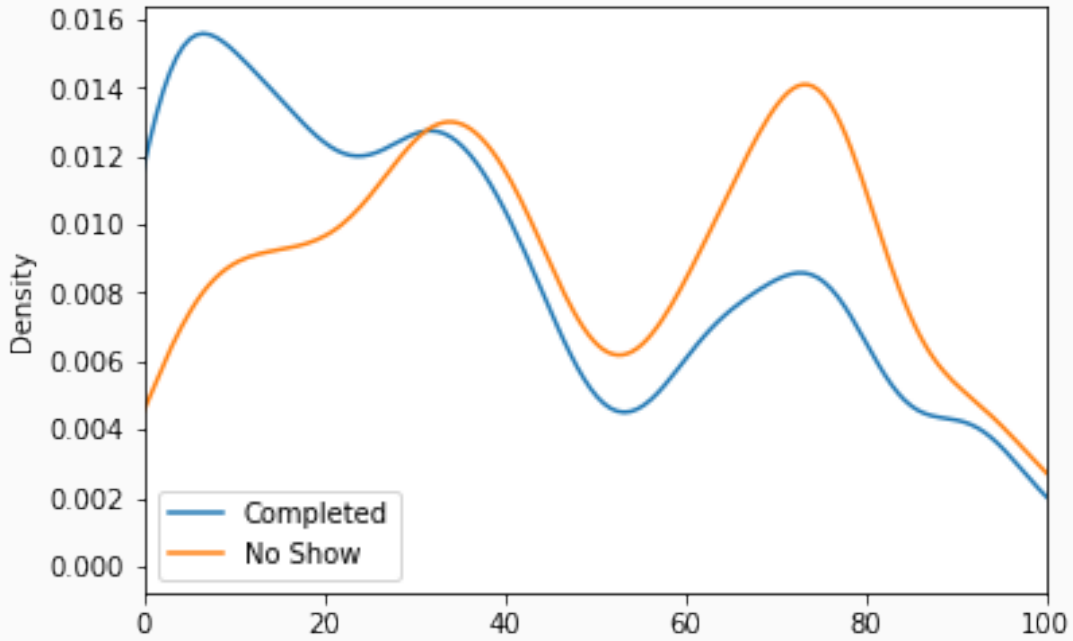
Figure 8: Pair plot with groups indicated



Visualization Toolbox

Bivariate exploration with target

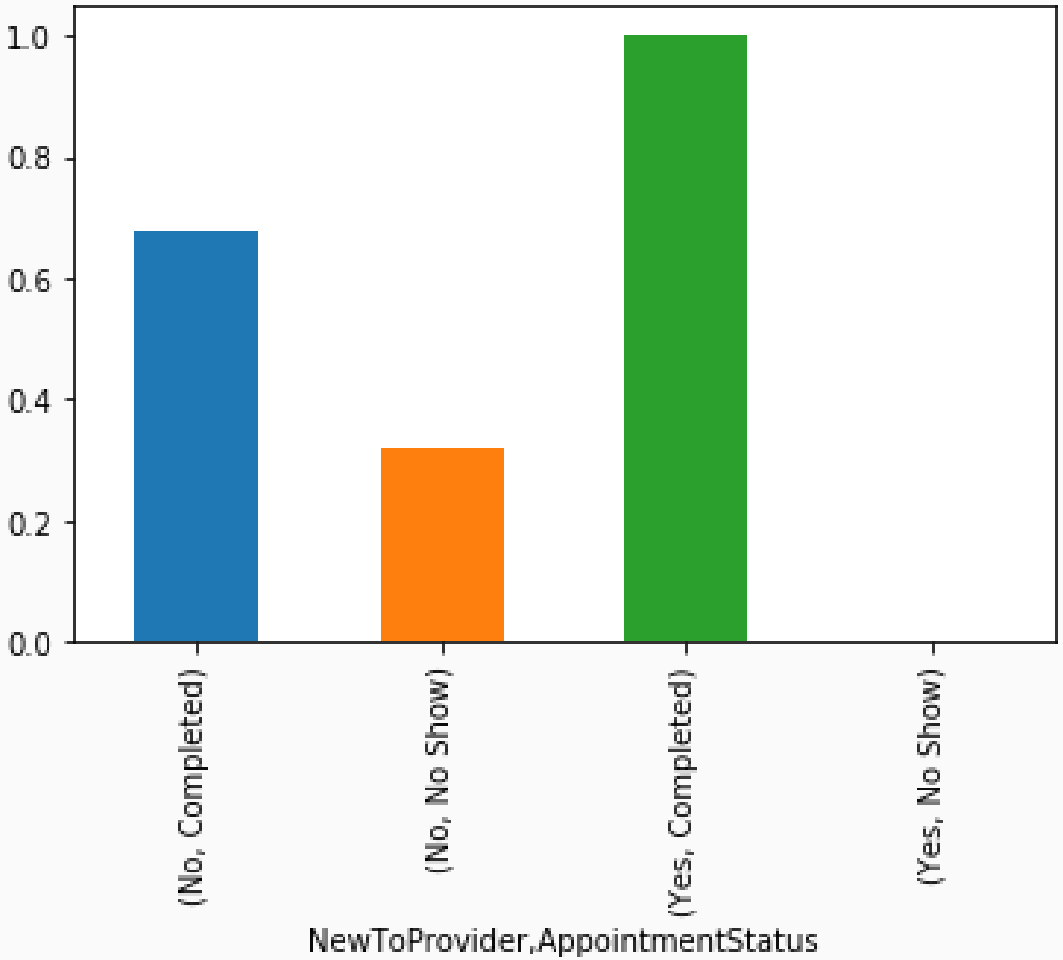
Figure 9: KDE Plot of variable with target groups



Visualization Toolbox

Bivariate exploration with target

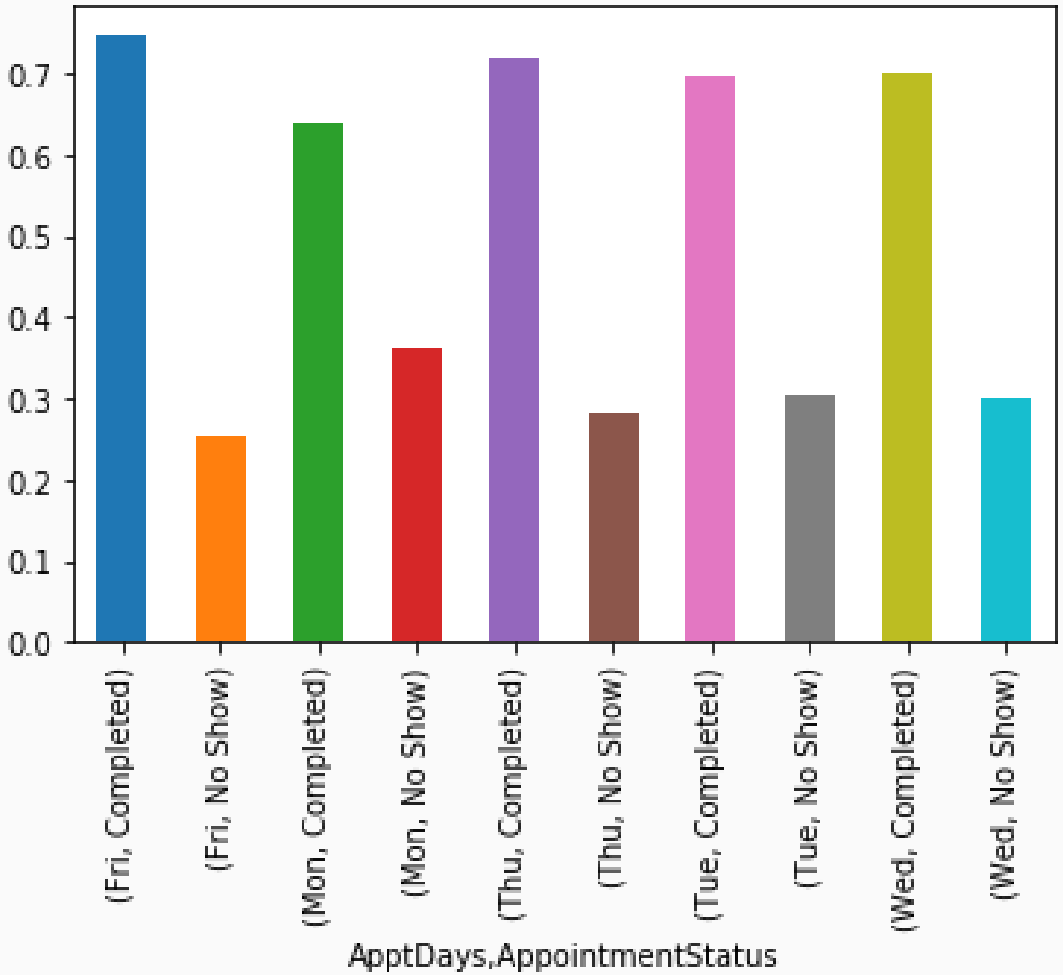
Figure 10: Count plot with target



Visualization Toolbox

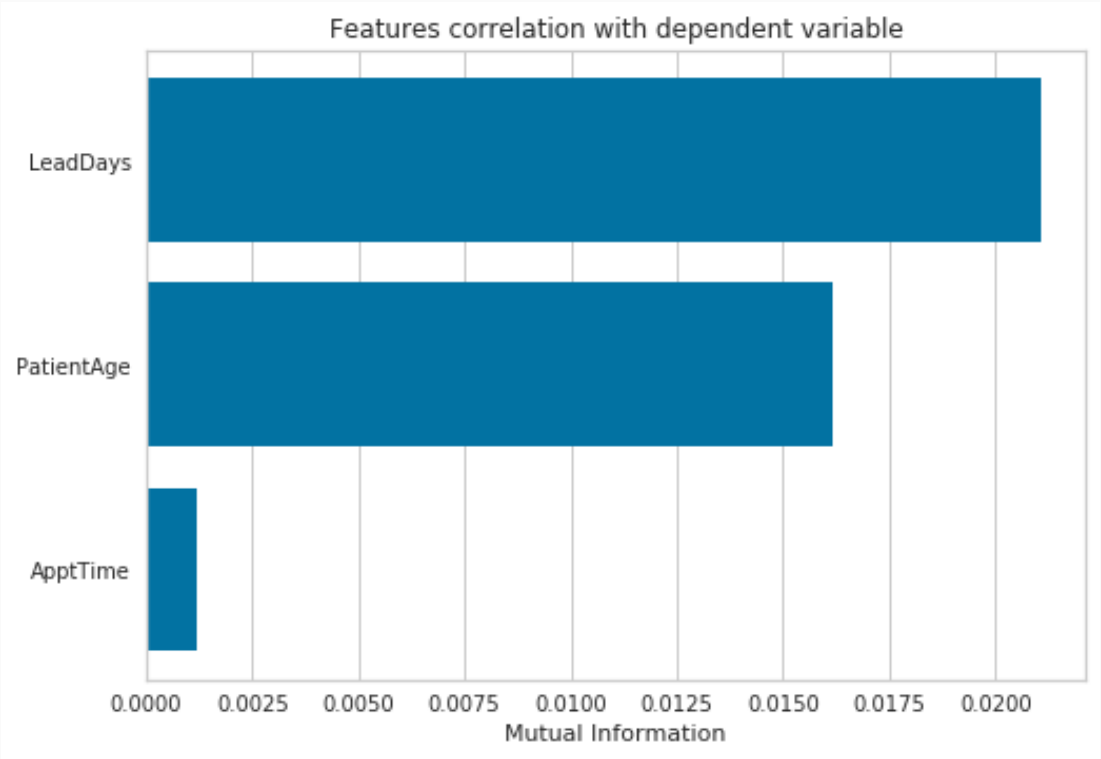
Bivariate exploration with target

Figure 11: Counts of discreet variables



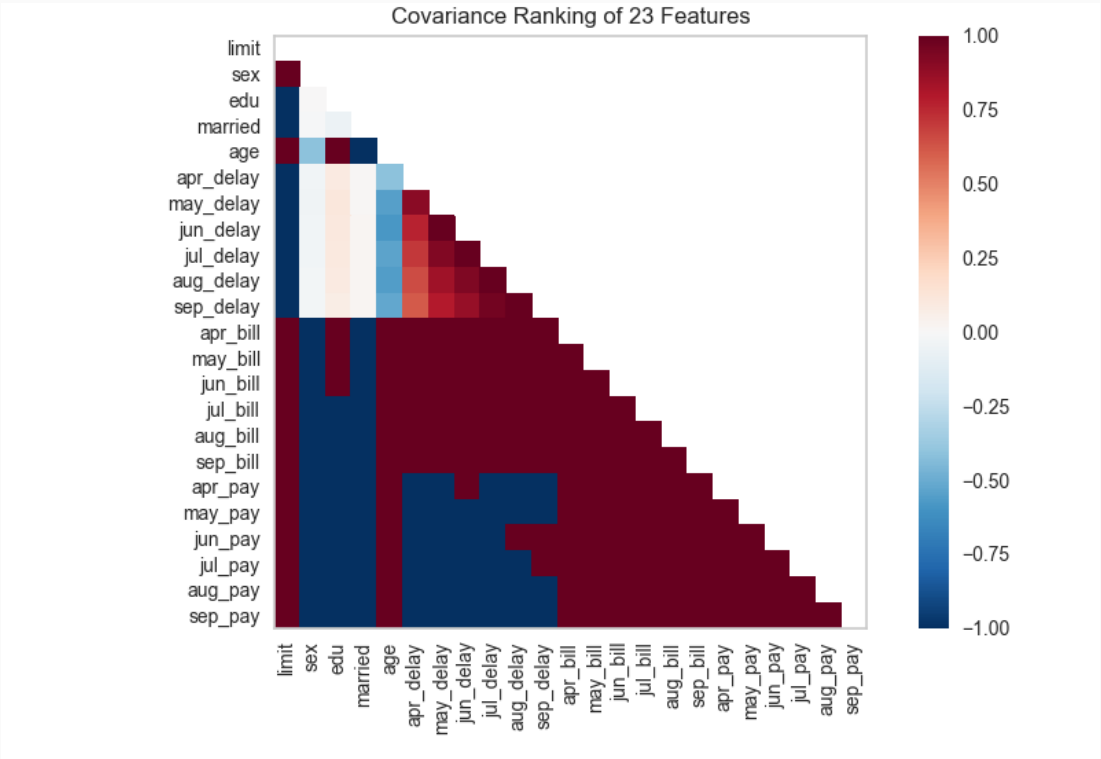
Feature Selection

Figure 12: Relationship between the variables and the target



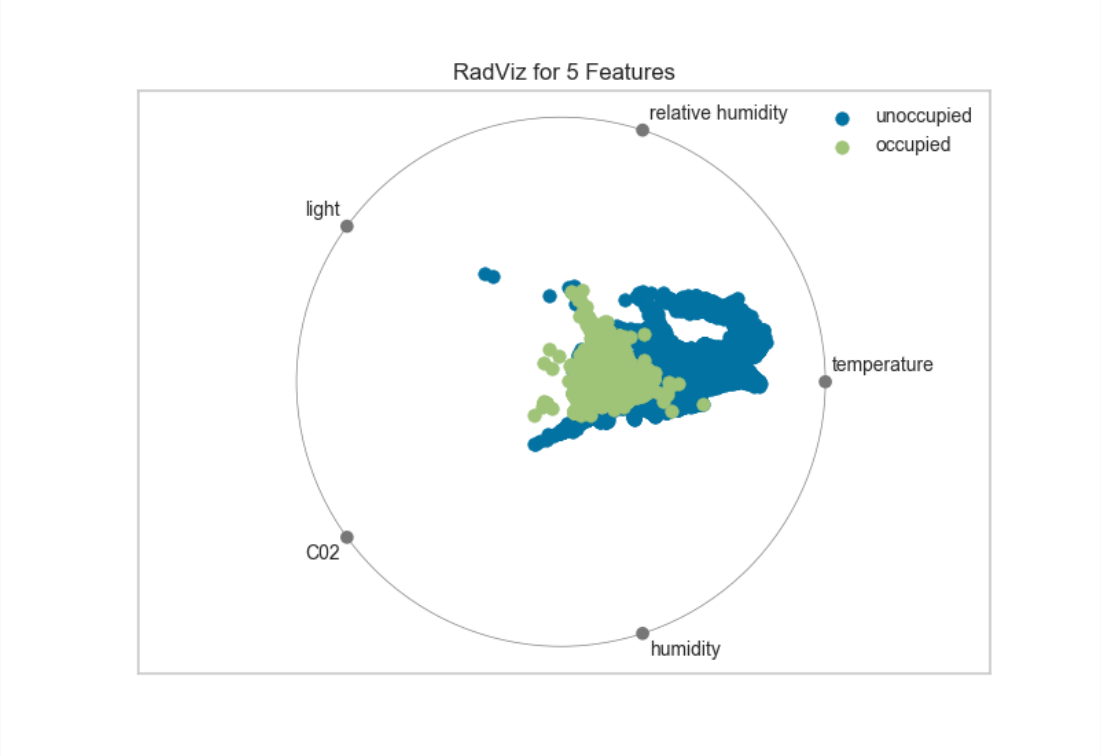
Other Visualizations

Figure 13: Covariance between variables



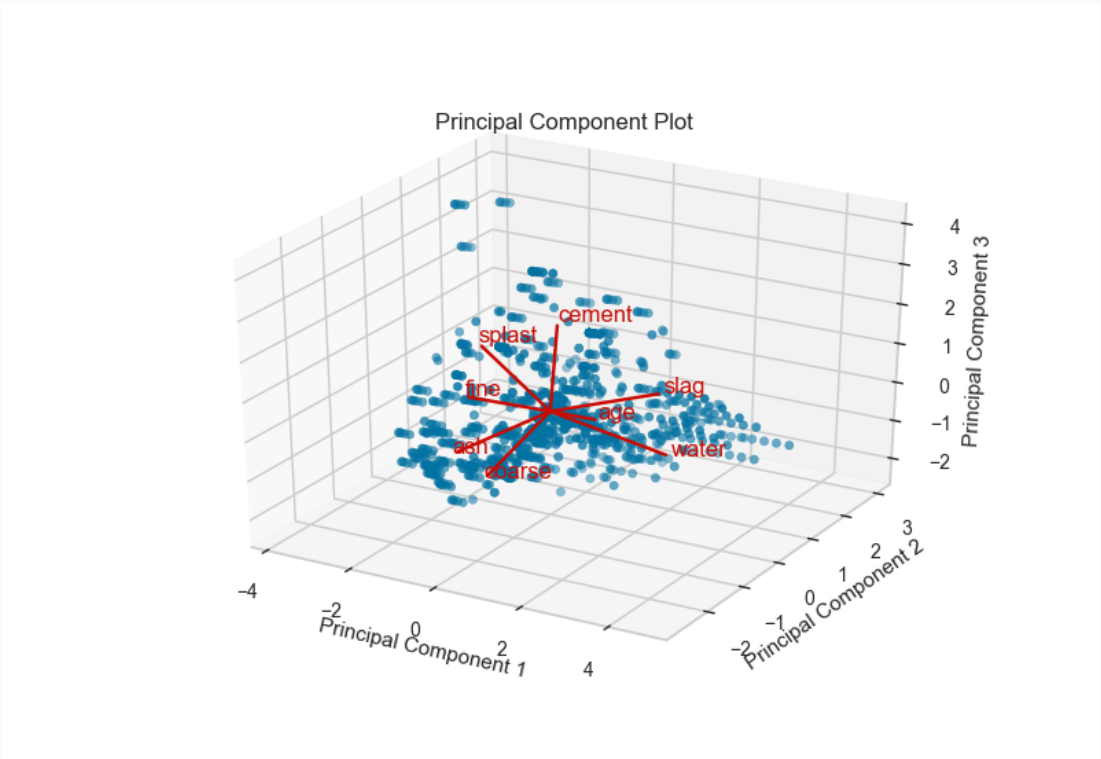
Other Visualizations

Figure 14: Radial plot to determine variable separability



Other Visualizations

Figure 15: PCA with axis plotted



Other Visualizations

Figure 16: Parallel coordinate plot