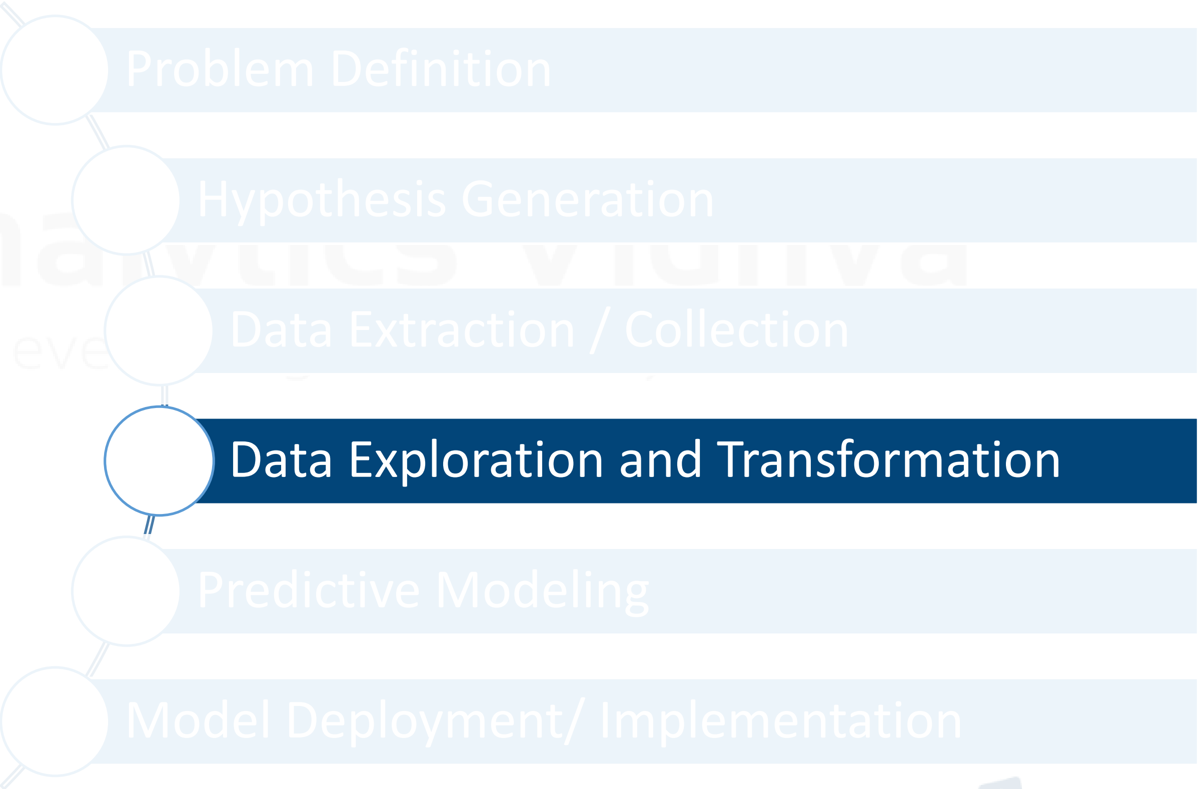
Predictive Modeling

Stages of Predictive Modeling

We can broadly divide the model building life cycle in six stages:



Problem Definition

Hypothesis Generation

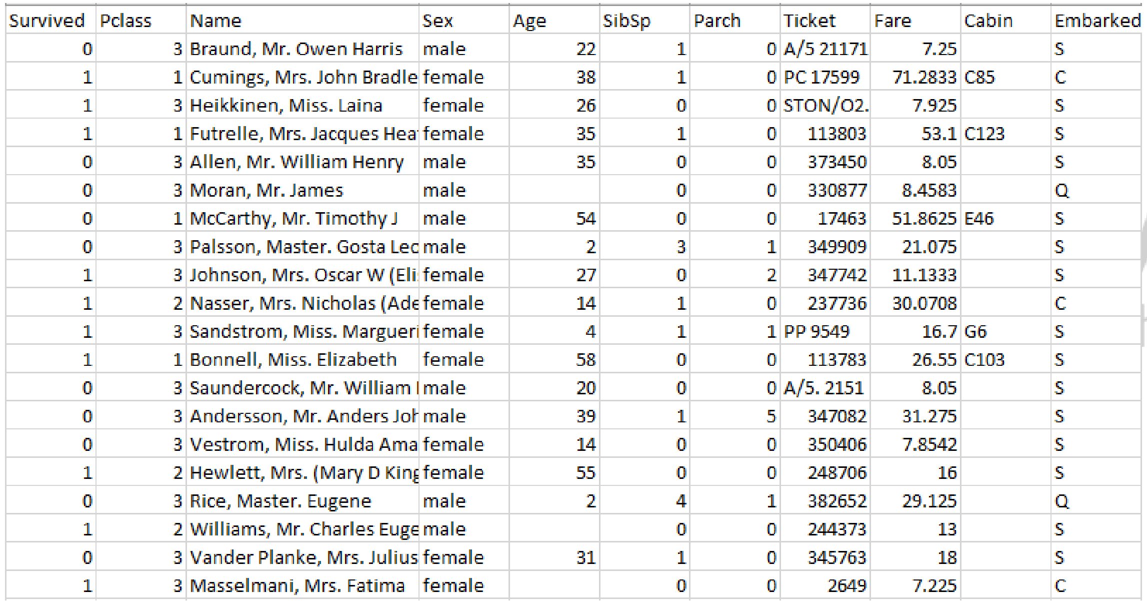
Data Extraction / Collection

Data Exploration and Transformation

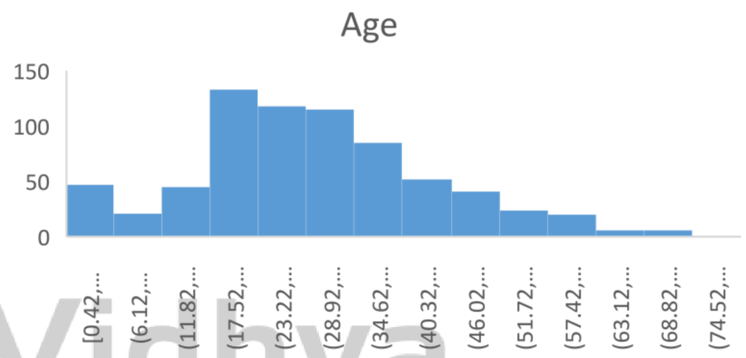
Predictive Modeling

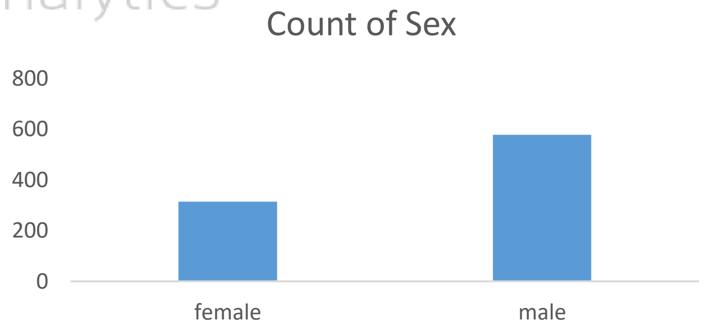
Model Deployment/ Implementation

Data Exploration



Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
1	1	Cumings, Mrs. John Bradle	female	38	1	0	PC 17599	71.2833	C85	C
1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
1	1	Futrelle, Mrs. Jacques Hea	female	35	1	0	113803	53.1	C123	S
0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
0	3	Palsson, Master. Gosta Lec	male	2	3	1	349909	21.075		S
1	3	Johnson, Mrs. Oscar W (Eli	female	27	0	2	347742	11.1333		S
1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
1	3	Sandstrom, Miss. Margueri	female	4	1	1	PP 9549	16.7	G6	S
1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
0	3	Saundercock, Mr. William I	male	20	0	0	A/5. 2151	8.05		S
0	3	Andersson, Mr. Anders Joh	male	39	1	5	347082	31.275		S
0	3	Vestrom, Miss. Hulda Ama	female	14	0	0	350406	7.8542		S
1	2	Hewlett, Mrs. (Mary D King	female	55	0	0	248706	16		S
0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
1	2	Williams, Mr. Charles Euge	male		0	0	244373	13		S
0	3	Vander Planke, Mrs. Julius	female	31	1	0	345763	18		S
1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C





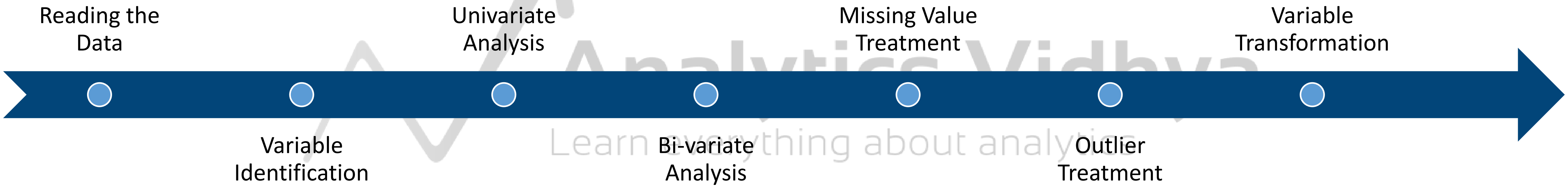
Data Exploration

Good Analyst vs Bad Analyst :

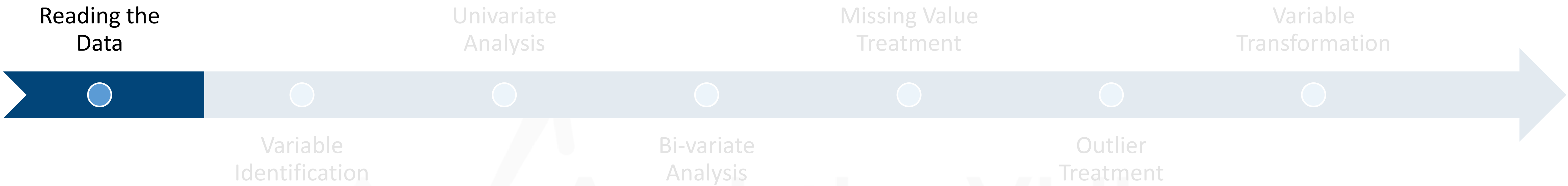
Good Analyst – Knows his/her data well.

Bad Analyst – Relies on tools and libraries

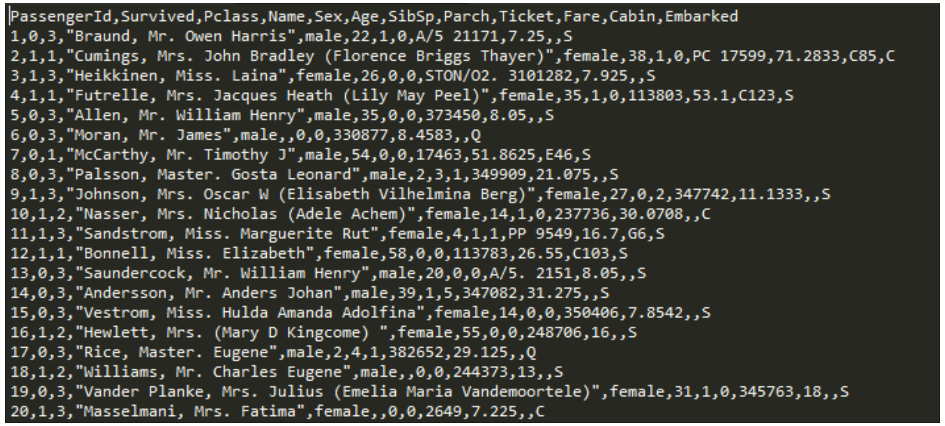
Data Exploration/ Transformation



Steps for Data Exploration

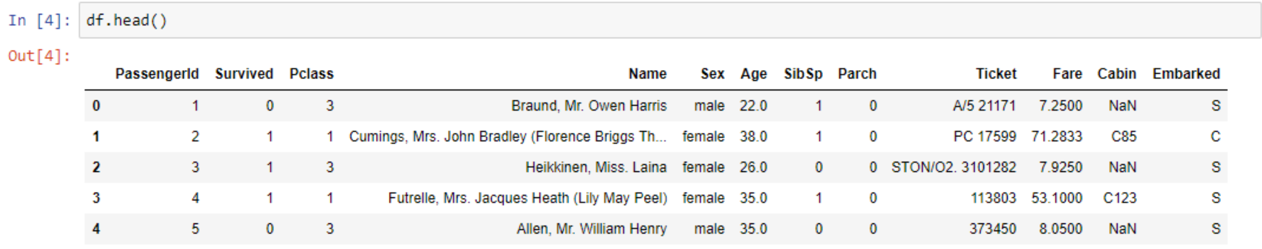


Raw csv



```
PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,S
4,1,1,"Futrelle, Mrs. Jacques Heath (Lily May Peel)",female,35,1,0,113803,53.1,C123,S
5,0,3,"Allen, Mr. William Henry",male,35,0,0,373450,8.05,,S
6,0,3,"Moran, Mr. James",male,,0,0,330877,8.4583,,Q
7,0,1,"McCarthy, Mr. Timothy J",male,54,0,0,17463,51.8625,E46,S
8,0,3,"Palsson, Master. Gosta Leonard",male,2,3,1,349909,21.075,,S
9,1,3,"Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)",female,27,0,2,347742,11.1333,,S
10,1,2,"Nasser, Mrs. Nicholas (Adele Achem)",female,14,1,0,237736,30.0708,,C
11,1,3,"Sandstrom, Miss. Marguerite Rut",female,4,1,1,PP 9549,16.7,G6,S
12,1,1,"Bonnell, Miss. Elizabeth",female,58,0,0,113783,26.55,C103,S
13,0,3,"Saundercock, Mr. William Henry",male,20,0,0,A/5. 2151,8.05,,S
14,0,3,"Andersson, Mr. Anders Johan",male,39,1,5,347082,31.275,,S
15,0,3,"Vestrom, Miss. Hulda Amanda Adolfina",female,14,0,0,350406,7.8542,,S
16,1,2,"Hewlett, Mrs. (Mary D Kingcome) ",female,55,0,0,248706,16,,S
17,0,3,"Rice, Master. Eugene",male,2,4,1,382652,29.125,,Q
18,1,2,"Williams, Mr. Charles Eugene",male,,0,0,244373,13,,S
19,0,3,"Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)",female,31,1,0,345763,18,,S
20,1,3,"Masselmani, Mrs. Fatima",female,,0,0,2649,7.225,,C
```

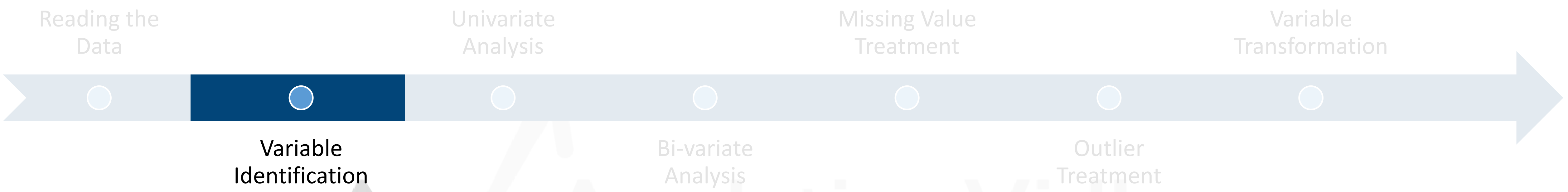
Pandas Dataframe- Python



```
In [4]: df.head()
Out[4]:
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Steps for Data Exploration



Reading the Data

Variable Identification

Univariate Analysis

Bi-variate Analysis

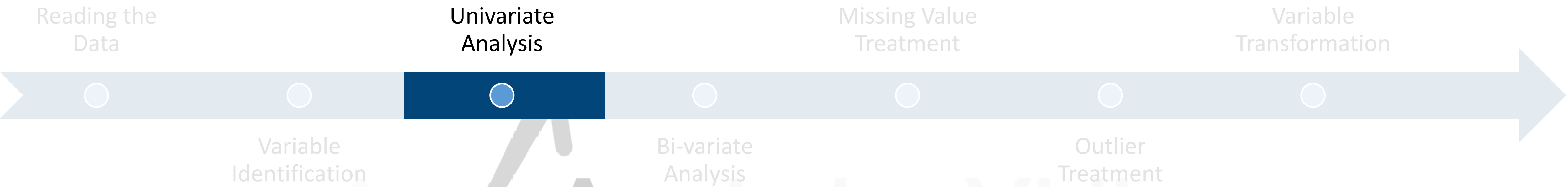
Missing Value Treatment

Outlier Treatment

Variable Transformation

Type of Variable	Data Type	Variable Category
Predictor -Pclass -Sex -Age Target - Survived	Character -Name -Ticket -Sex Numeric -Age -Pclass	Categorical -Sex -Cabin -Embarked Continuous -SibSp -Parch

Steps for Data Exploration



Reading the Data

Variable Identification

Univariate Analysis

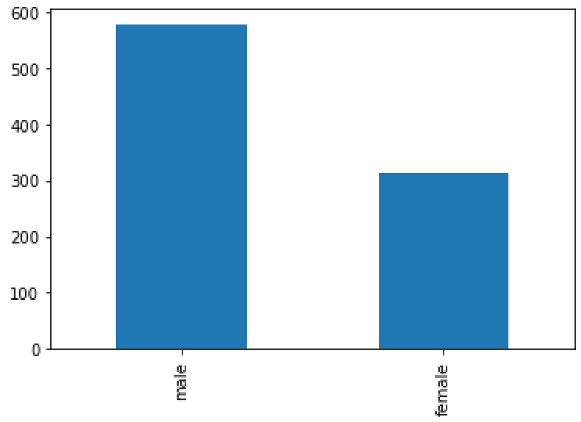
Bi-variate Analysis

Missing Value Treatment

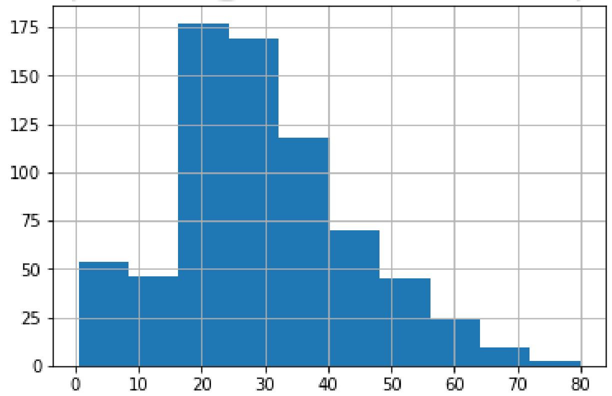
Outlier Treatment

Variable Transformation

We analyze variables one by one

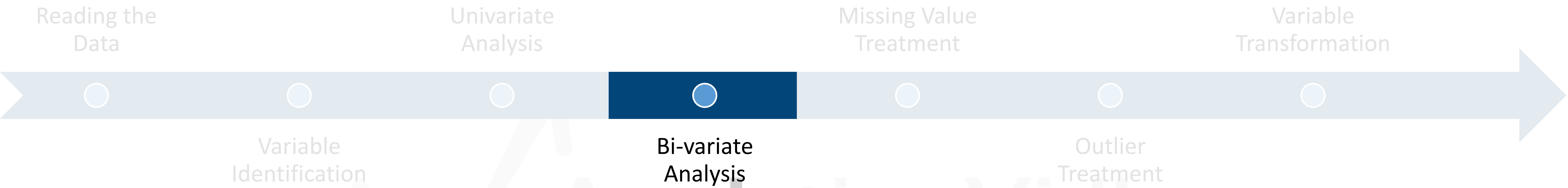


Bar plot of count of gender



Histogram of age

Steps for Data Exploration



Reading the Data

Variable Identification

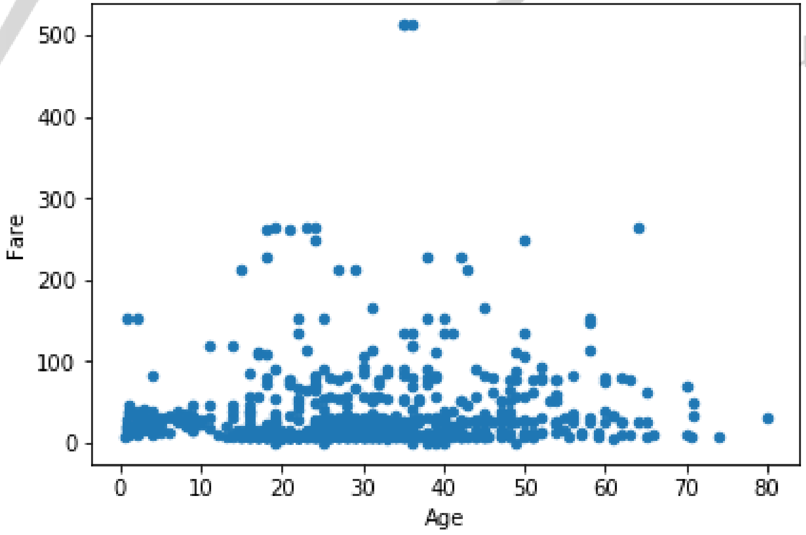
Univariate Analysis

Bi-variate Analysis

Missing Value Treatment

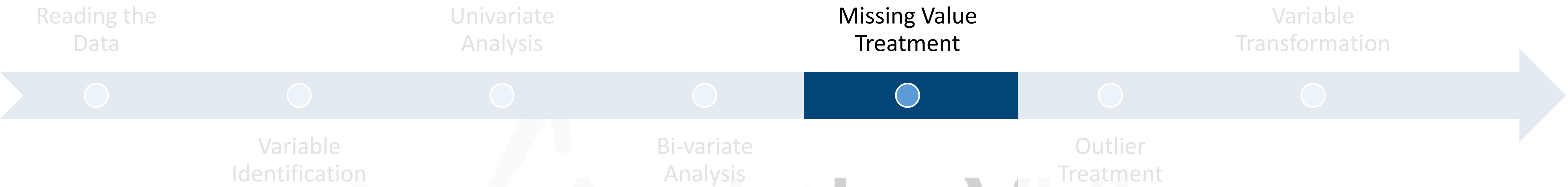
Outlier Treatment

Variable Transformation



Analytics Vidhya
Learn everything about analytics

Steps for Data Exploration



Reading the Data

Variable Identification

Univariate Analysis

Bi-variate Analysis

Missing Value Treatment

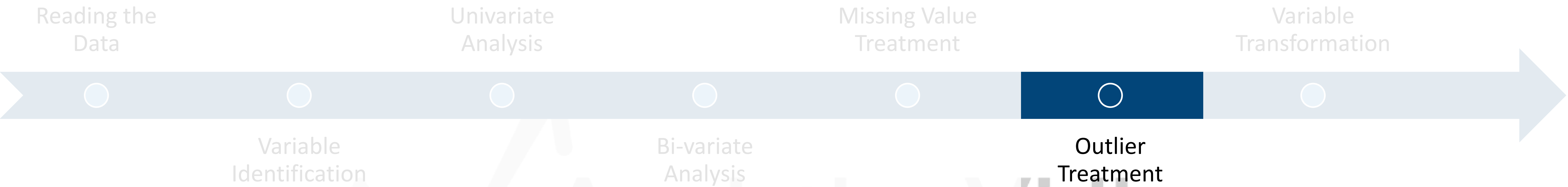
Outlier Treatment

Variable Transformation

Find out and treat missing values in the dataset using -

1. Mean
2. Mode
3. Median

Steps for Data Exploration



Reading the Data

Variable Identification

Univariate Analysis

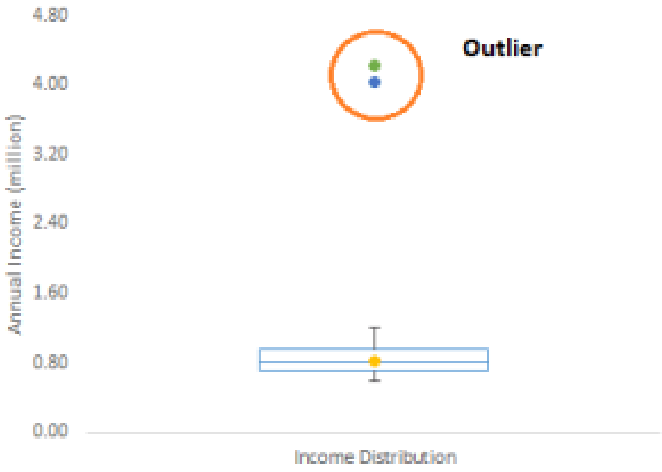
Bi-variate Analysis

Missing Value Treatment

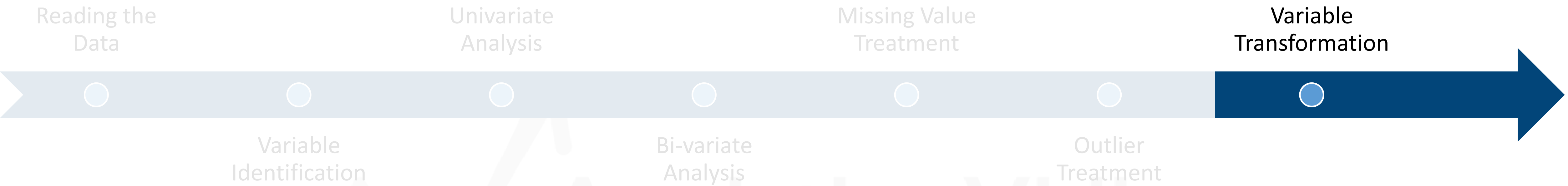
Outlier Treatment

Variable Transformation

Outlier is an observation that appears far away and diverges from an overall pattern in a sample.



Steps for Data Exploration



Reading the Data

Variable Identification

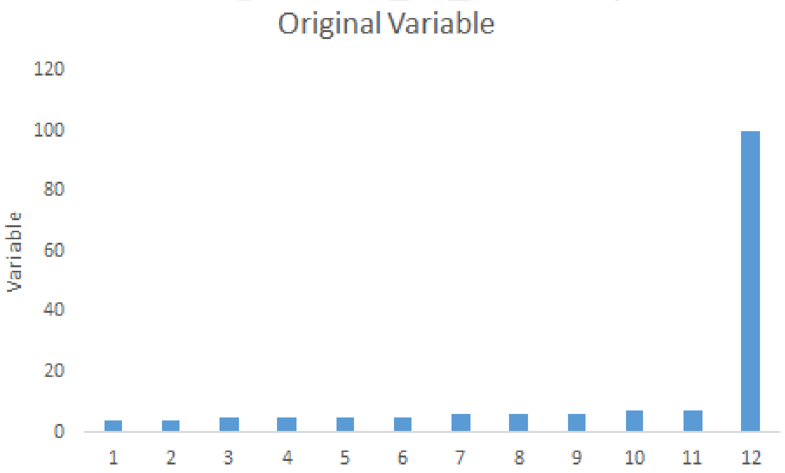
Univariate Analysis

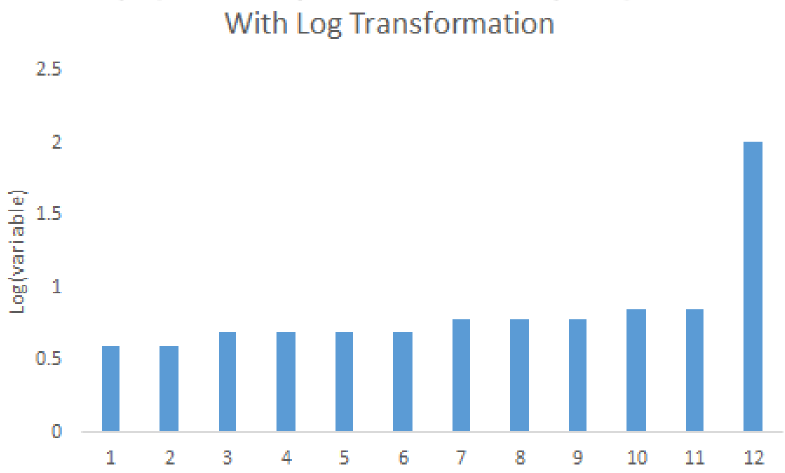
Bi-variate Analysis

Missing Value Treatment

Outlier Treatment

Variable Transformation





Analytics Vidhya
Learn everything about analytics

Thank You