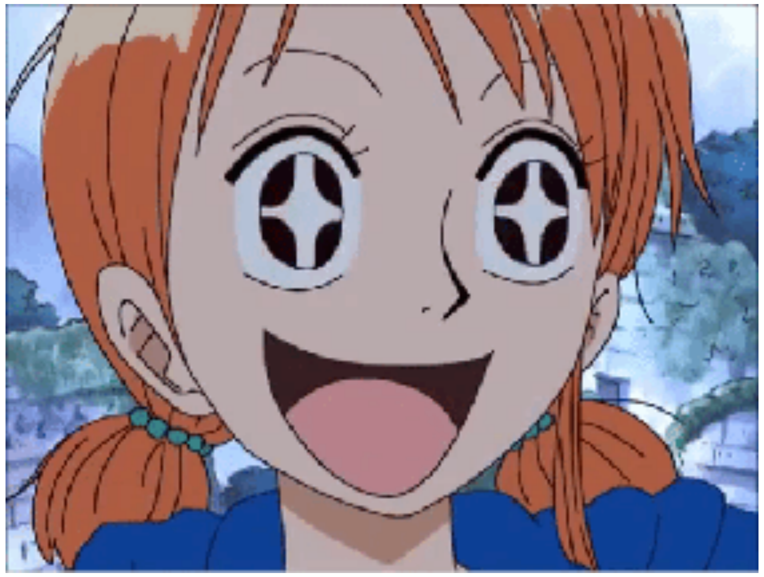
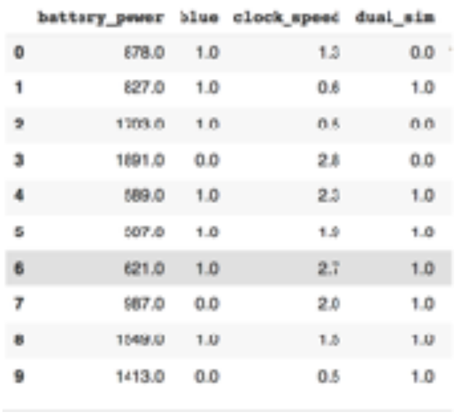
Feature Engineering

Why feature engineering?

We face lot of problems with our dataset which makes our lives just

Let's see those problems and find solutions





	battery_power	blue	clock_speed	dual_sim
0	678.0	1.0	1.3	0.0
1	627.0	1.0	0.6	1.0
2	1703.0	1.0	0.5	0.0
3	1891.0	0.0	2.8	0.0
4	589.0	1.0	2.3	1.0
5	507.0	1.0	1.9	1.0
6	621.0	1.0	2.7	1.0
7	987.0	0.0	2.0	1.0
8	1049.0	1.0	1.5	1.0
9	1413.0	0.0	0.5	1.0

1- missed value data
2- diversity data
3- outlier in database

1 Scatter plotted the column

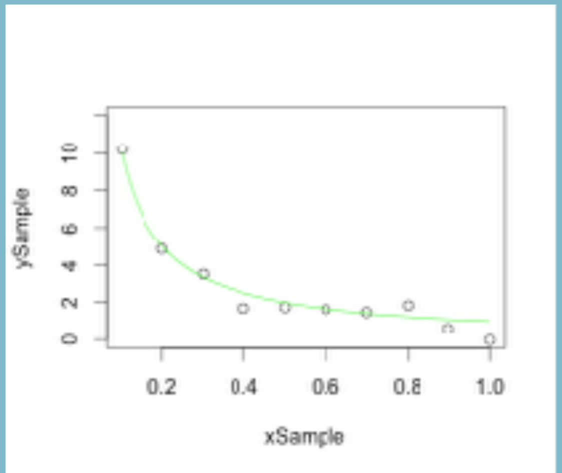


Fig 1

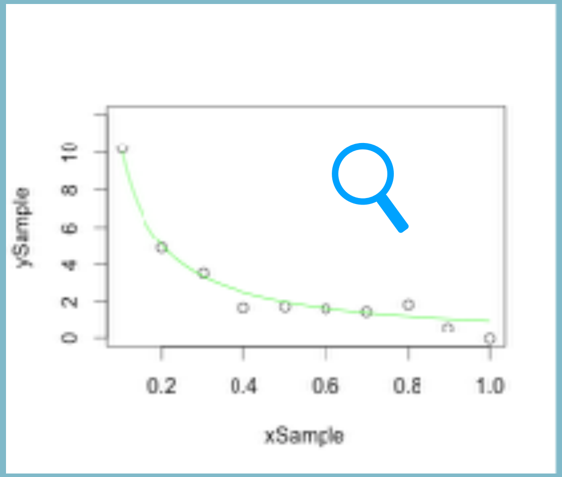


Fig 2

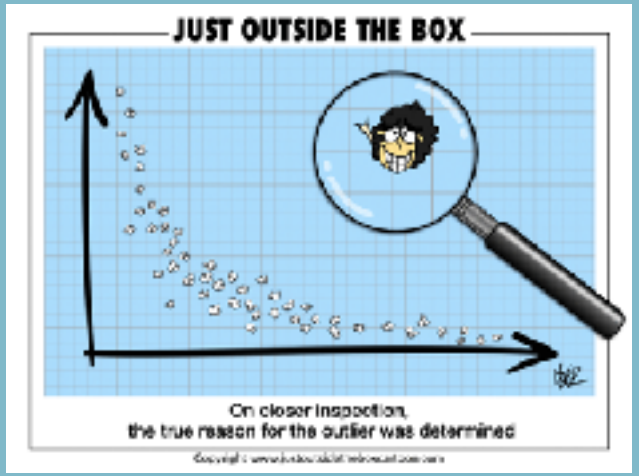


Fig 3

Outlier

Problems that we might face with such datasets are:

1. **How to find outliers?**
2. **How to deal with these outliers?**

2 Histogram the column

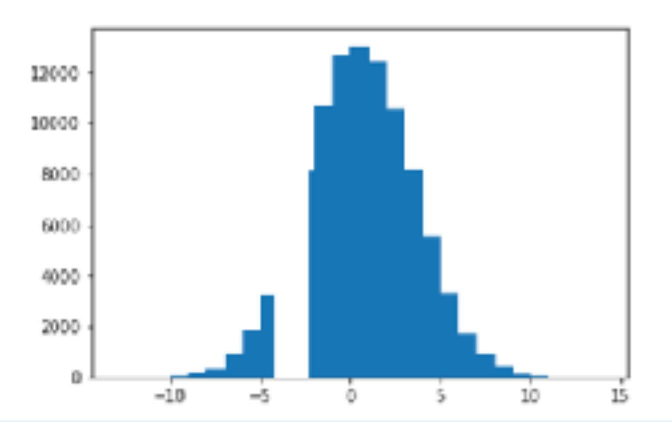


Fig 4

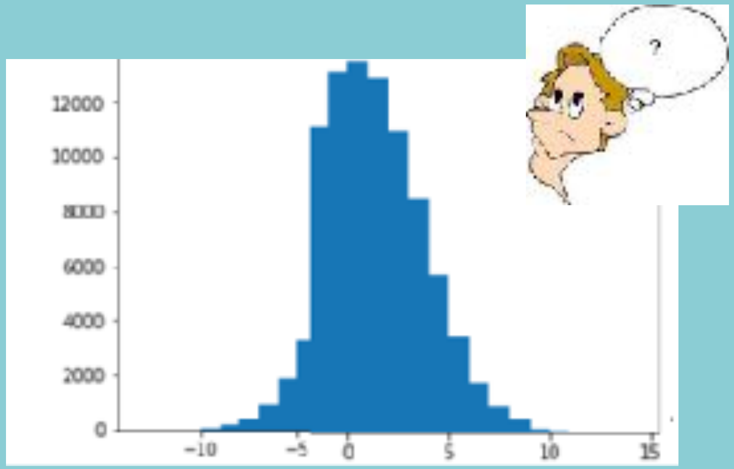


Fig 5

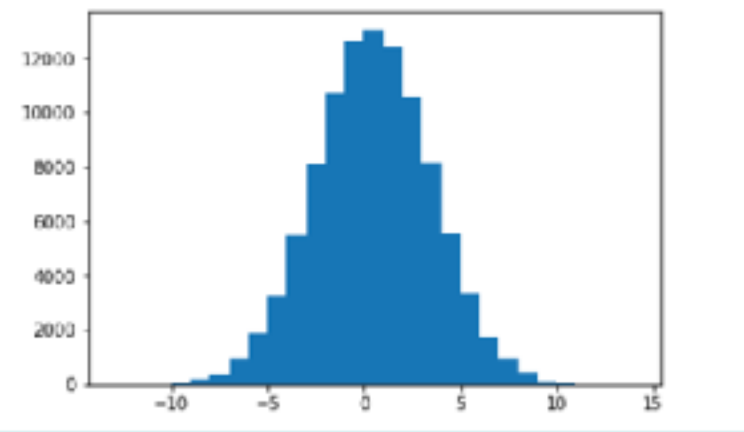


Fig 6

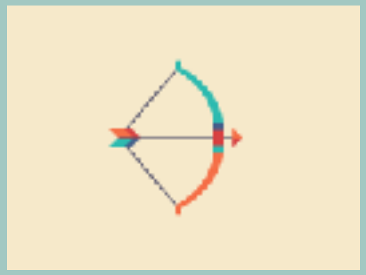
Missing data

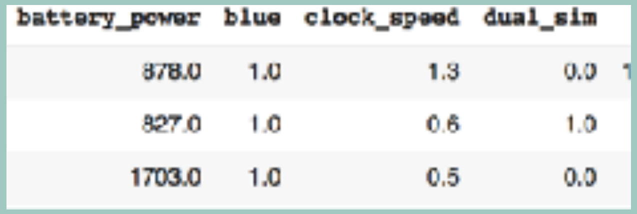
Problems that we might face with such datasets are:

1. **How to find missing data?**
2. **How to deal with these missing data?**

3 Diversity in data

Fig 1





battery_power	blue	clock_speed	dual_sim
878.0	1.0	1.3	0.0
827.0	1.0	0.6	1.0
1703.0	1.0	0.5	0.0

Fig 2

Different types of variables

Problems that we might face with such datasets are:

1. **What are the different types of variables/data?**
2. **How to deal with different types of variables?**

4 Impure Data / Feature selection

Fig 3

Fig 4

Fig 5

Irrelevant data

Problems that we might face with such datasets are:

1. **How to identify irrelevant data?**
2. **How to deal with irrelevant data?**

Feature engineering is a process of handling:

outliers

Missing data

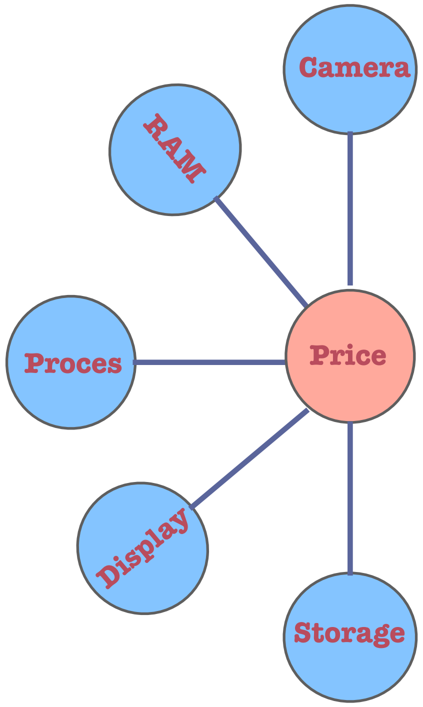
Converting between different data types
And removing irrelevant data

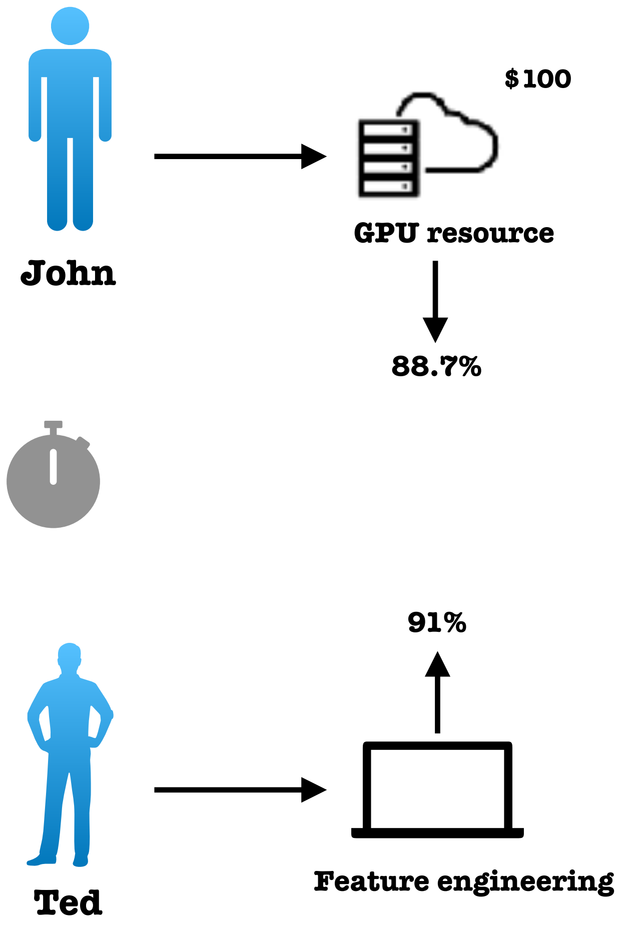
in order to make data more beautiful or meaningful to use.

What is feature engineering?

Feature engineering is a process of cleaning, transforming or covering between different data types in order to make data more relevant

Power of feature engineering

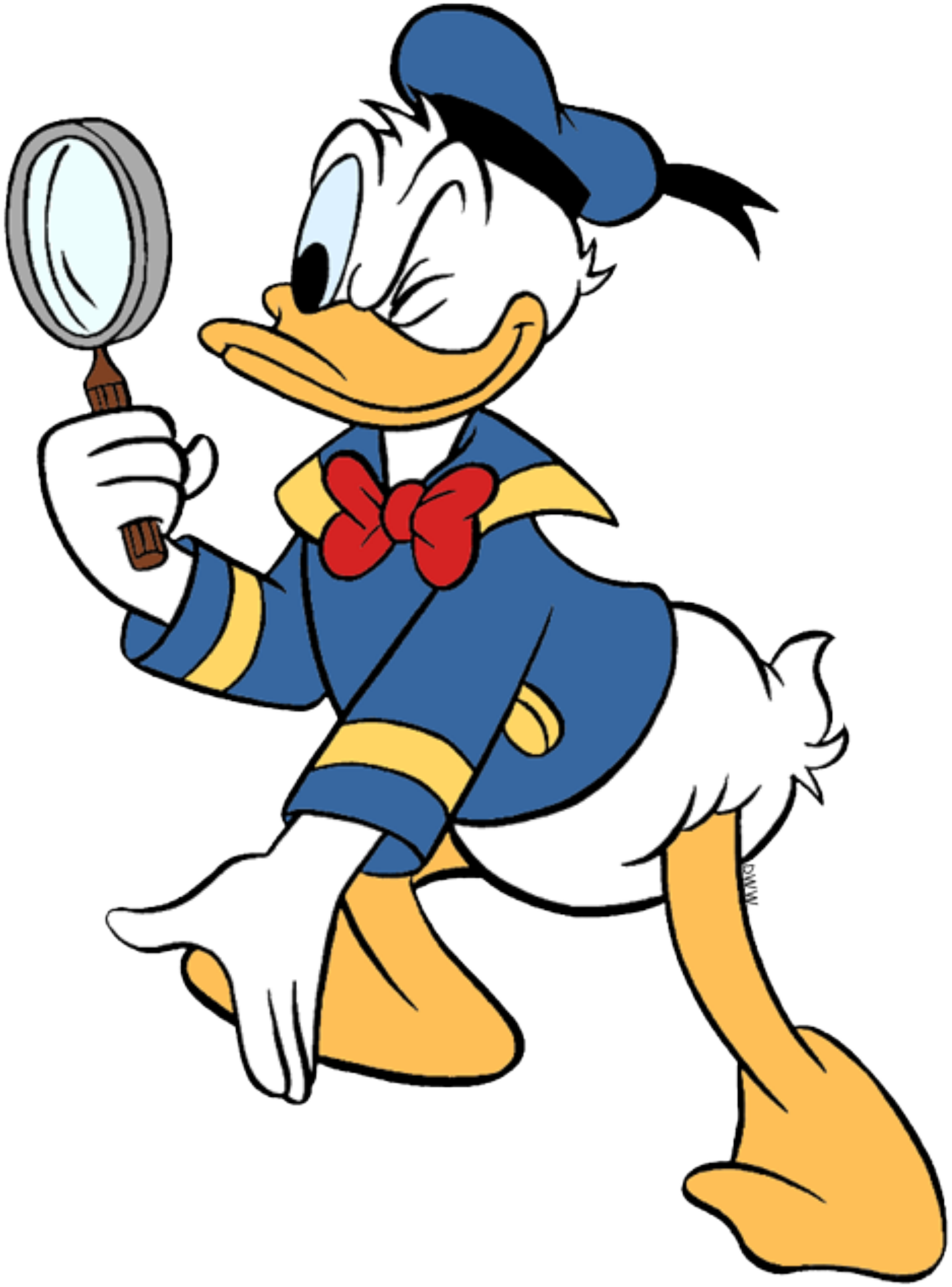




Vivo is a famous mobile company started buy Shen in 2009. When Shen started his company he wanted to give tough fight to big companies like Apple,Samsung etc.
He did not know how to estimate price of mobiles his company manufactured. To solve this problem he collects sales data of mobile phones of various companies.
Shen wanted to find out some relation between features of a mobile phone(eg:- RAM,Internal Memory etc) and its selling price.

Let's see how his team figured out the solution for this problem.

Understanding your data (1)



2. Feature engineering

1 Outlier

Problems that we might face with such datasets are:

1. How to find outliers?
2. How to deal with these outliers?

2 Missing data

Problems that we might face with such datasets are:

1. How to find missing data?
2. How to deal with these missing data?

3 Different types of variables

Problems that we might face with such datasets are:

1. What are the different types of variables/data?
2. How to deal with different types of variables?

4 Irrelevant data

Problems that we might face with such datasets are:

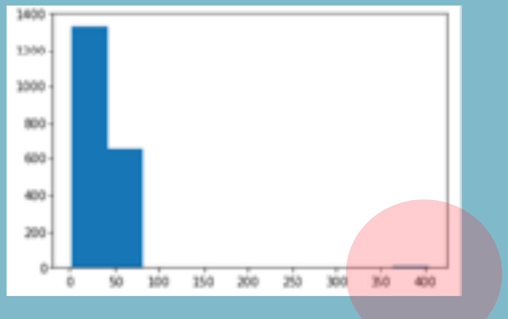
1. How to identify irrelevant data?
2. How to deal with irrelevant data?

2.1 Outlier

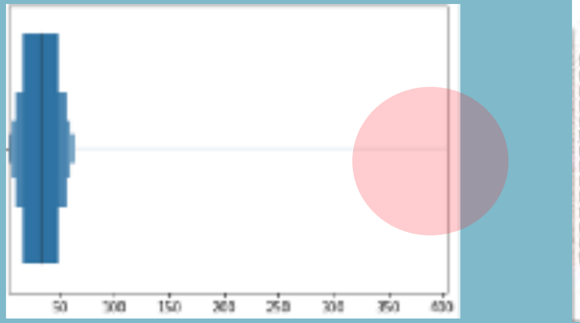
Problems that we might face with such datasets are:

1. **How to find outliers?**
2. **How to deal with these outliers?**

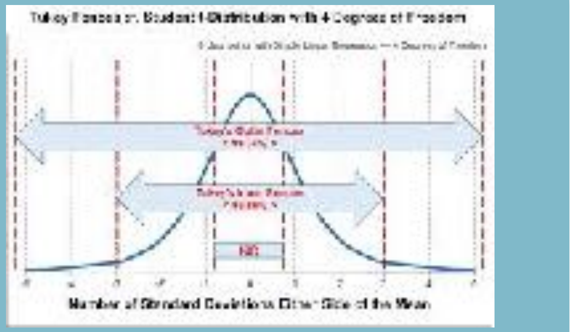
1. How to find outliers?



Histogram



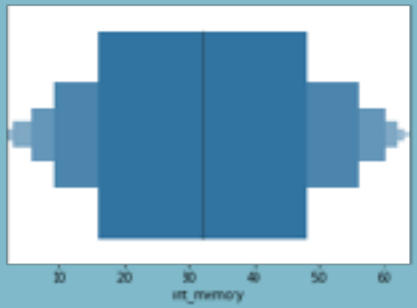
Boxplot



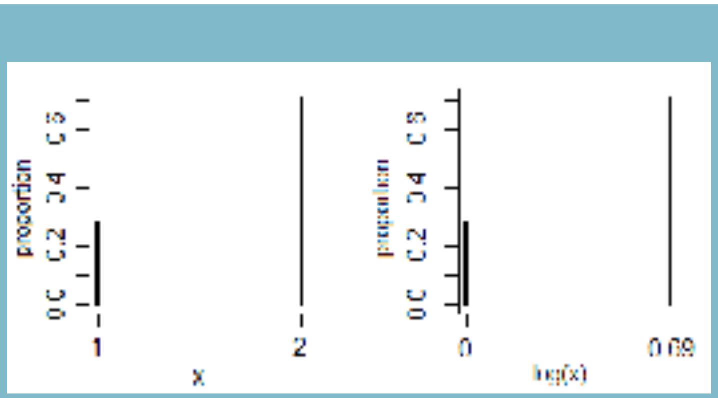
Tukey IQR

Finding

2. How to deal with these outliers?



Removing



Log transforming

Dealing

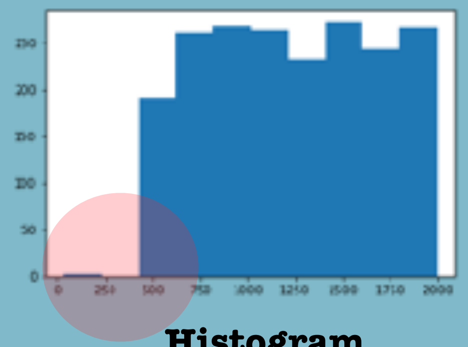
Front view
Top view
Log Trans

2.2 Missing data

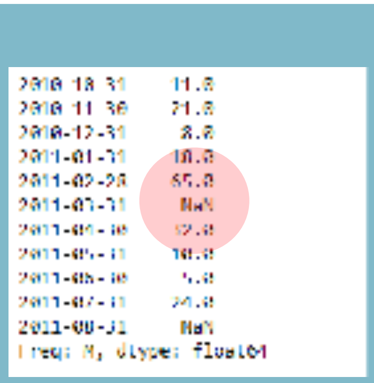
Problems that we might face with such datasets are:

1. **How to find missing data?**
2. **How to deal with these missing data?**

1. How to find missing data?



Histogram

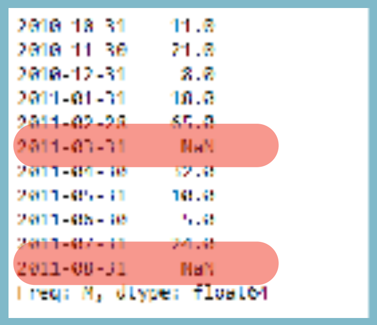


```
2010-10-31    11.0
2010-11-30    21.0
2010-12-31     8.0
2011-01-31    10.0
2011-02-28    65.0
2011-03-31     NaN
2011-04-30    12.0
2011-05-31    10.0
2011-06-30     5.0
2011-07-31    24.0
2011-08-31     NaN
Freq: M, dtype: float64
```

DataFrame functions

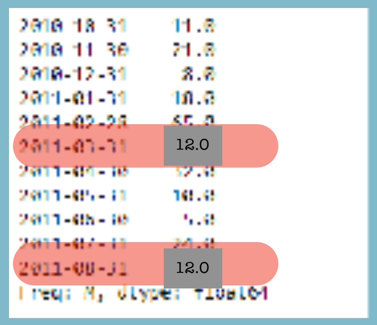
Finding

2. How to deal with missing data?



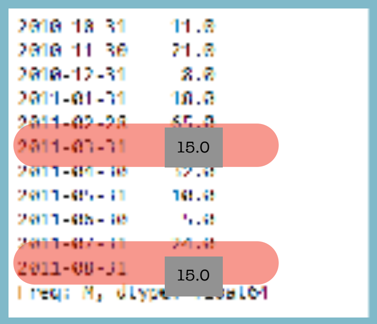
```
2010-10-31    11.0
2010-11-30    21.0
2010-12-31     8.0
2011-01-31    10.0
2011-02-28    65.0
2011-03-31     NaN
2011-04-30    12.0
2011-05-31    10.0
2011-06-30     5.0
2011-07-31    24.0
2011-08-31     NaN
Freq: M, dtype: float64
```

Removing



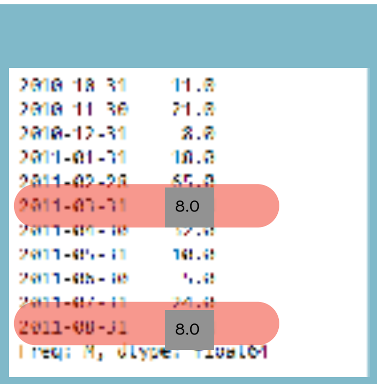
```
2010-10-31    11.0
2010-11-30    21.0
2010-12-31     8.0
2011-01-31    10.0
2011-02-28    65.0
2011-03-31    12.0
2011-04-30    12.0
2011-05-31    10.0
2011-06-30     5.0
2011-07-31    24.0
2011-08-31    12.0
Freq: M, dtype: float64
```

Replace with Mean



```
2010-10-31    11.0
2010-11-30    21.0
2010-12-31     8.0
2011-01-31    10.0
2011-02-28    65.0
2011-03-31    15.0
2011-04-30    12.0
2011-05-31    10.0
2011-06-30     5.0
2011-07-31    24.0
2011-08-31    15.0
Freq: M, dtype: float64
```

Replace with Median



```
2010-10-31    11.0
2010-11-30    21.0
2010-12-31     8.0
2011-01-31    10.0
2011-02-28    65.0
2011-03-31     8.0
2011-04-30    12.0
2011-05-31    10.0
2011-06-30     5.0
2011-07-31    24.0
2011-08-31     8.0
Freq: M, dtype: float64
```

Replace with Most freq

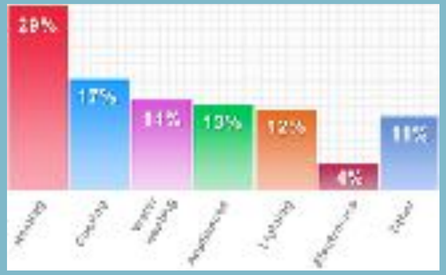
Dealing

2.3 Different types of variables

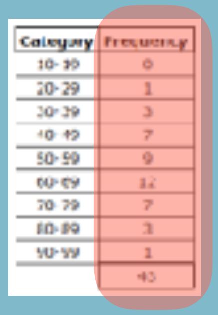
Problems that we might face with such datasets are:

1. **What are the different types of variables/data?**
2. **How to deal with different types of variables?**

1. What are the different types of variables/data?



Categorical data



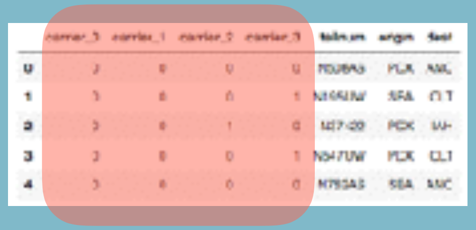
Category	Frequency
10-19	0
20-29	1
30-39	3
40-49	7
50-59	9
60-69	12
70-79	7
80-89	3
90-99	1
	43

Numerical data

Text data

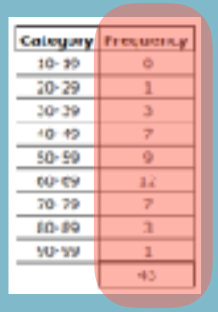
Types

2. How to deal with different types of variables?



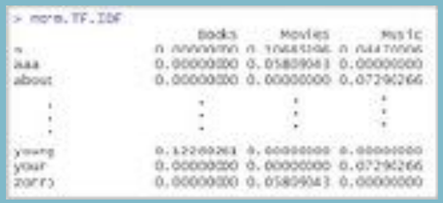
	carrier_0	carrier_1	carrier_2	carrier_3	tailnum	origin	dest
0	[illegible]	0	0	0	[illegible]	PDX	ANC
1	[illegible]	0	0	1	[illegible]	SEA	CLT
2	[illegible]	0	1	0	[illegible]	PDX	IAH
3	[illegible]	0	0	1	[illegible]	PDX	CLT
4	[illegible]	0	0	0	[illegible]	SEA	ANC

Encode categorical features



Category	Frequency
10-19	0
20-29	1
30-39	3
40-49	7
50-59	9
60-69	12
70-79	7
80-89	3
90-99	1
	43

Numerical data



```
> norm.TF.IDF
                Books     Movies      Music
a          0.00000000 [illegible] [illegible]
aaa        0.00000000 0.05809043 0.00000000
about      0.00000000 0.00000000 0.07296266
   .            .          .          .
   .            .          .          .
   .            .          .          .
young      [illegible] 0.00000000 0.00000000
your       0.00000000 0.00000000 0.07296266
zorro      0.00000000 0.05809043 0.00000000
```

Text to numerical features
TFIDF

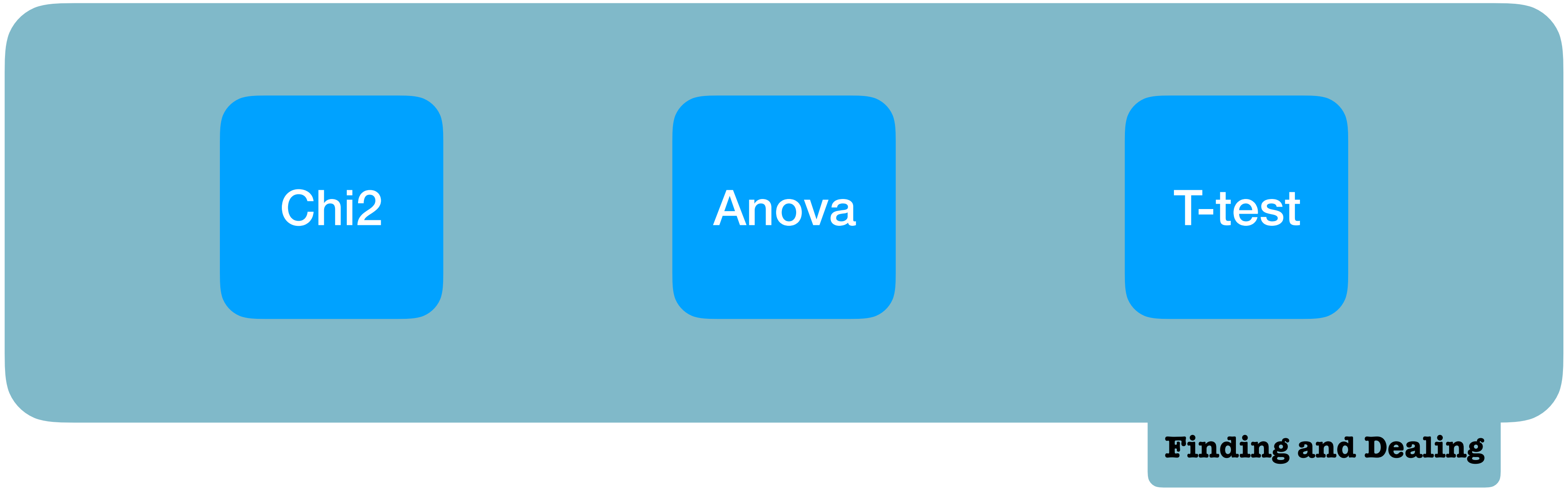
Dealing

2.4 Irrelevant data

Problems that we might face with such datasets are:

1. **How to identify irrelevant data?**
2. **How to deal with irrelevant data?**

1. **How to identify irrelevant data?**
2. **How to deal with irrelevant data?**

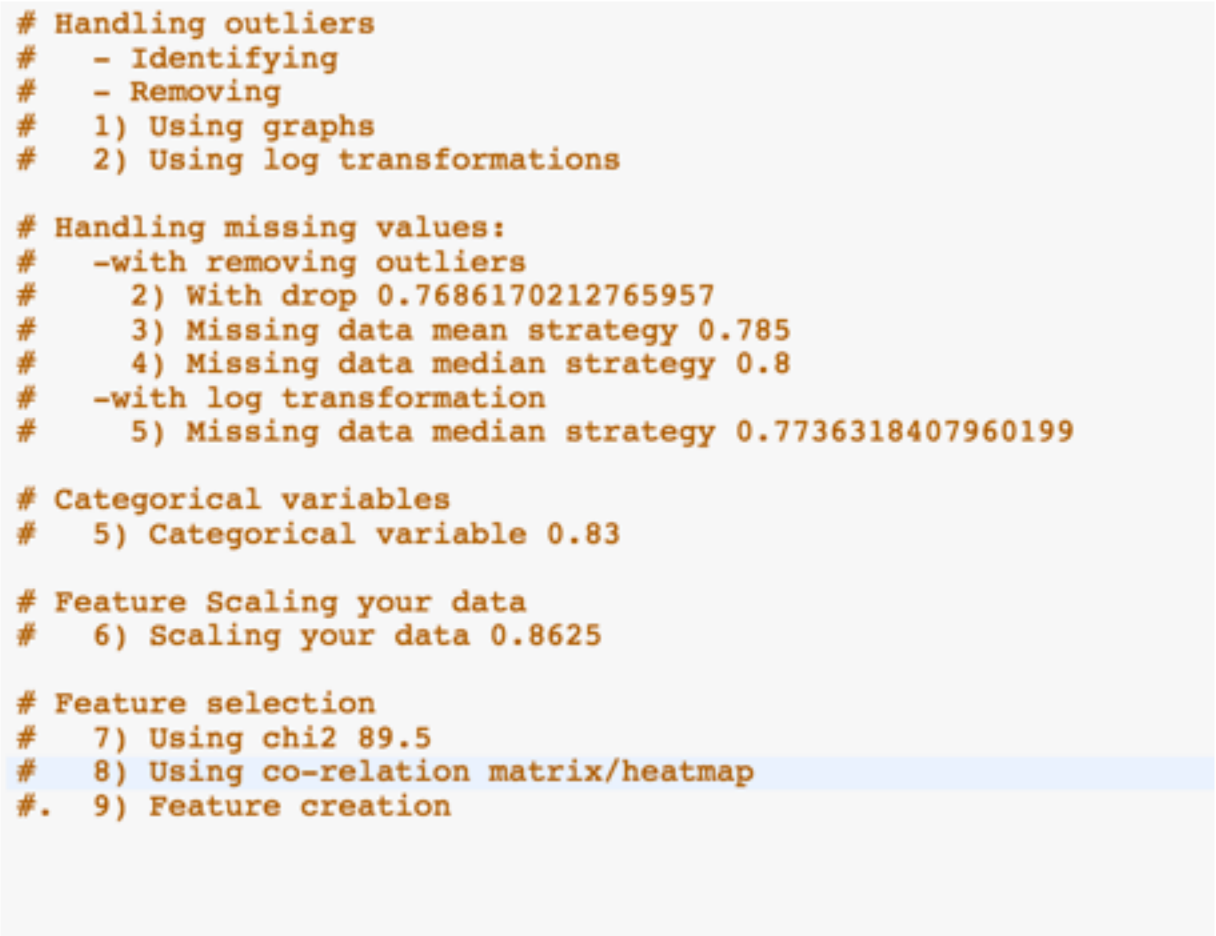


Chi2

Anova

T-test

Finding and Dealing



```
# Handling outliers
#   - Identifying
#   - Removing
#   1) Using graphs
#   2) Using log transformations

# Handling missing values:
#   -with removing outliers
#     2) With drop 0.7686170212765957
#     3) Missing data mean strategy 0.785
#     4) Missing data median strategy 0.8
#   -with log transformation
#     5) Missing data median strategy 0.7736318407960199

# Categorical variables
#   5) Categorical variable 0.83

# Feature Scaling your data
#   6) Scaling your data 0.8625

# Feature selection
#   7) Using chi2 89.5
#   8) Using co-relation matrix/heatmap
#.  9) Feature creation
```

Thank you