I

Introduction to Data Science

Data Science Academy

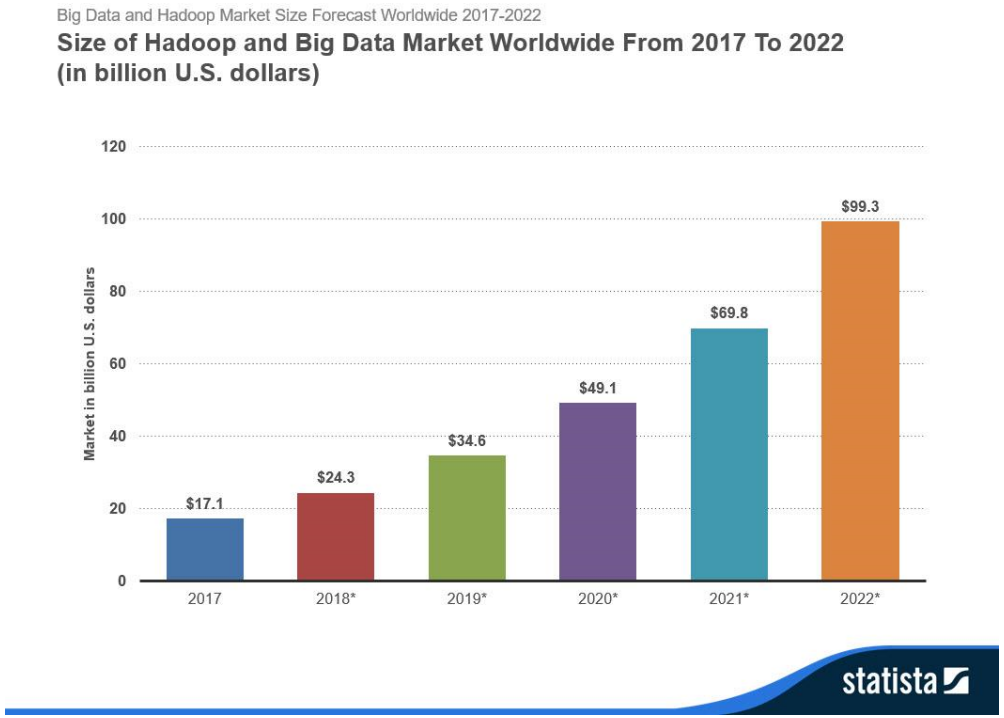
Data Science has been one of the biggest tech buzz word for the last 5-10 years!

- Data Science, Artificial Intelligence, Machine Learning, Big Data
- Those terms have been bouncing around every tech site and been heavily glamourized (even vilified) by the media!

Bütün hüquqlar qorunur.

The Big Data Industry is Growing Rapidly!

IDC Forecasts Revenues for Big Data and Business Analytics Solutions Will Reach $189.1 Billion This Year with Double-Digit Annual Growth Through 2022

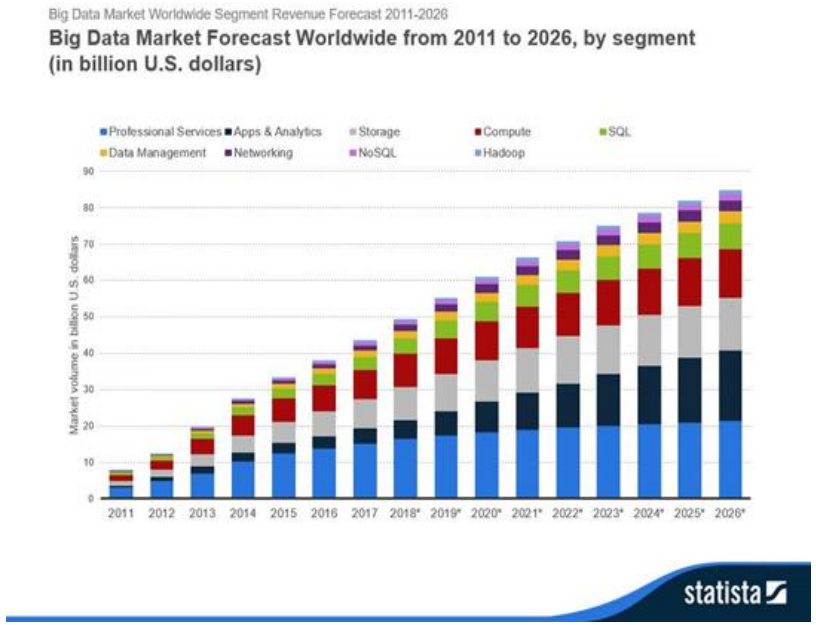


Bütün hüquqlar qorunur.

The Demand for Data Scientists is only going up!

Demand for data scientists is booming and will only increase

Fueled by big data and AI, demand for data science skills is growing exponentially, according to job sites. The supply of skilled applicants, however, is growing at a slower pace.



18,002 views | Oct 14, 2019, 07:15am

The Birth Of The Data Science Generation

Forbes Technology Council

Ike Kavas Forbes Councils Member
Forbes Technology Council COUNCIL POST | Paid Program
Innovation

Bütün hüquqlar qorunur.

Has this ever happened to you?

- You're thinking about something, maybe it's the new printer you wanted, or a skiing trip you were planning. Or perhaps thinking about starting a gym.
- Then BAM! You see an online Ad for the exact thing you wanted.

Gmail

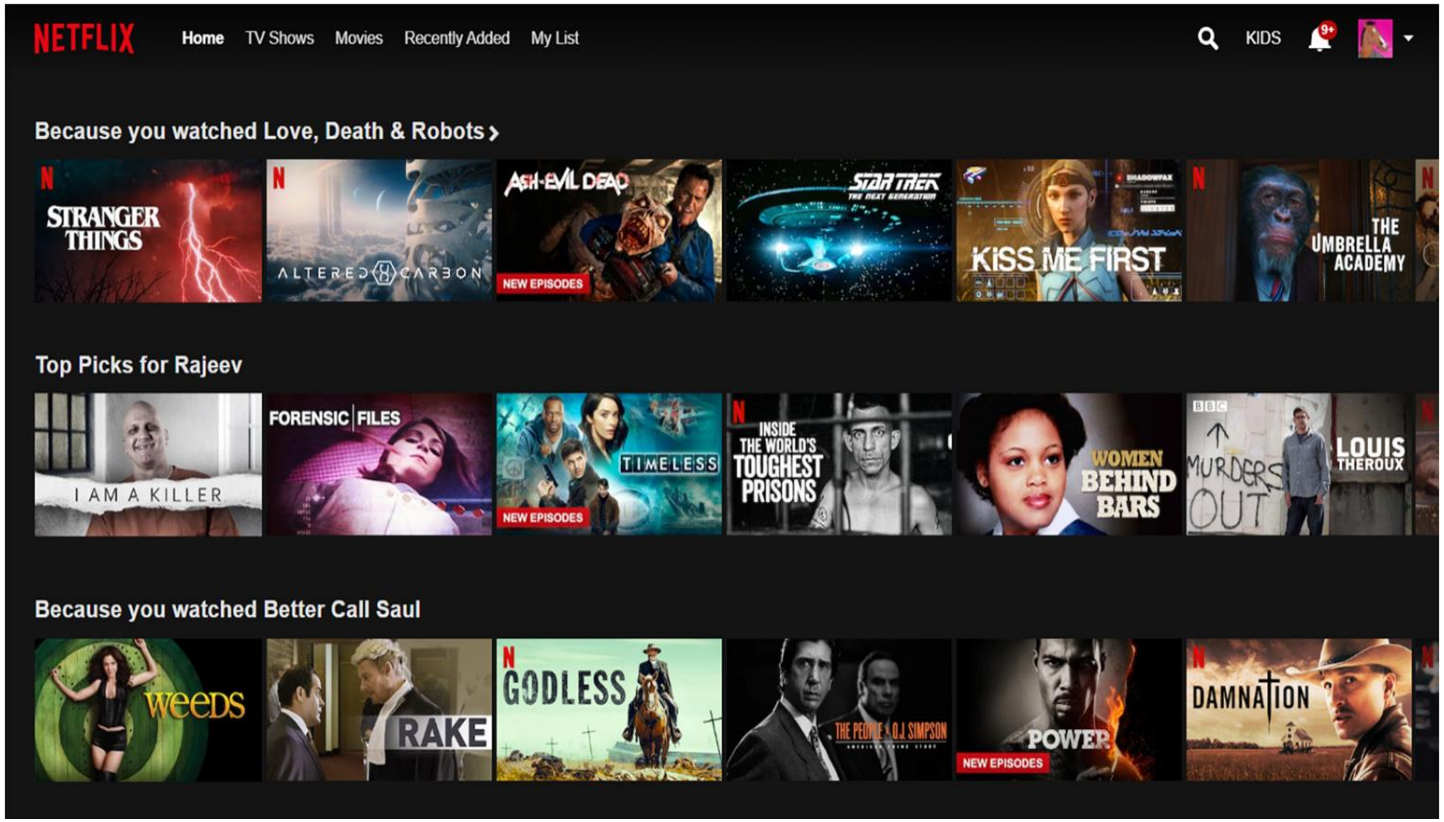
Bütün hüquqlar qorunur.

How do those online giants like Google and Facebook know you so well?

- Unfortunately, it's not magic
- It's DATA
- And it's everywhere (including yours)
- Let's take a look at 5 super interesting examples of how Data Driven companies are changing the business landscape

Bütün hüquqlar qorunur.

Netflix’s Recommendations



Bütün hüquqlar qorunur.

The Ubiquity of Data Opportunities

- Automated Recruitment of Employees
- Crypto, Forex & Stock Price Predictions, Credit Risk Scoring Models
- Health Analytics – Disease Prediction, finding cures etc.
- Computer Vision – Understanding what is being seen to build things like self driving cars and facial recognition
- Agriculture
- Manufacturing
- Creating Art and Music
- Self Driving cars and Robots
- Chat Bots
- And thousands more! There is no shortage on areas we can apply Data Science

Here's how data science helped Zoomcar capture 75% of Indian market

"Data lake and models built on Kafka helped us achieve enhanced customer experience at the best possible price," Arpit Agarwal, Director, Decision Science, Zoomcar, says.

Nikhar Aggarwal | ETCIO | October 31, 2019, 09:23 IST

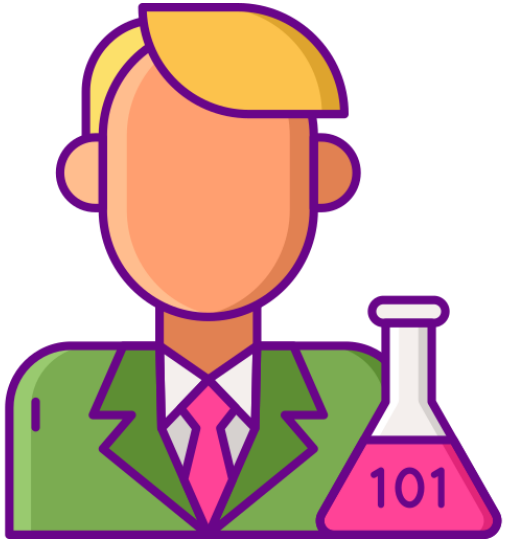
Bütün hüquqlar qorunur.

Data = Value = Better Decisions = More Profits!

Bütün hüquqlar qorunur.

Data Analysts vs. Data Scientists

No Machine Learning doesn't need to be an ace programmer



Uses Machine Learning and needs to be good at programming

Bütün hüquqlar qorunur.

How Data Scientists Approach Problems

Data Science Academy

What you're getting

- 1500+ slides
- 25+ cheat sheets
- Comprehensive Data Science theoretical and practical learning path
- ~50+ ipython notebooks
- ~80+ ipython notebooks
- ~80+ R Studio Scripts
- ~Professional Career services
- Mentorship
- GitHub Portfolio
- 24 Amazing Real World Projects
- Data Science Community

Bütün hüquqlar qorunur.

Course outline:

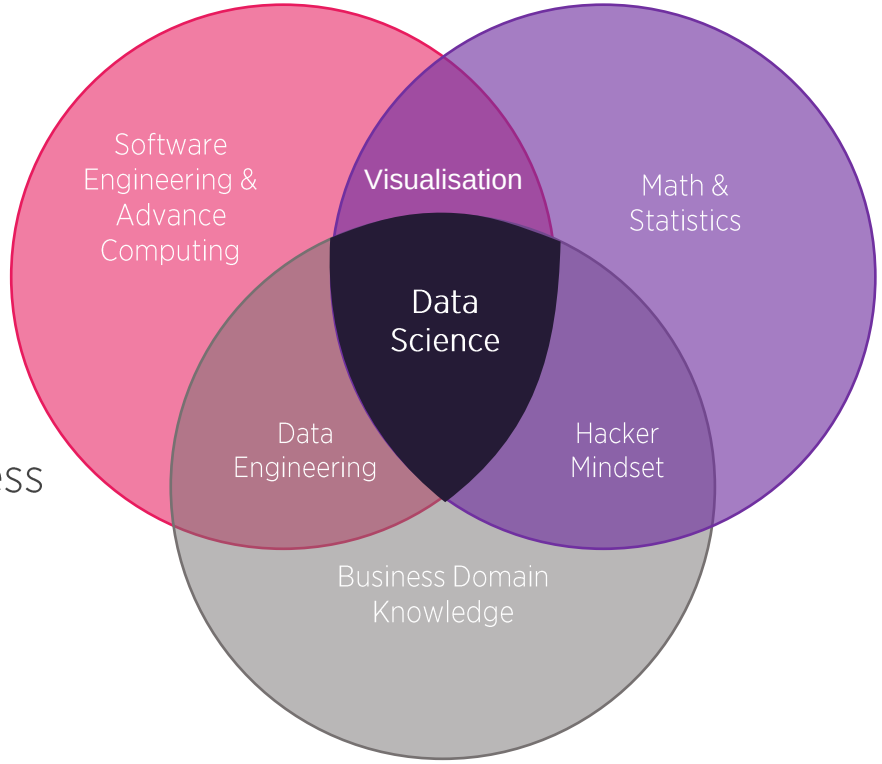
1. Data Science Divisions and Fields
2. What is Data Analytics & Statistics
3. Basic Statistical Concepts (Population, Sample)
4. Descriptive Statistics
5. Types of data
6. Levels of measurements
7. Descriptive statistics (Numerical Methods)
8. Measure of locations
9. Measure of variability
10. Normal Distribution
11. Z-score
12. Empiric Rule
13. Detecting Outliers
14. Skewness

Bütün hüquqlar qorunur.

What you'll be able to do after

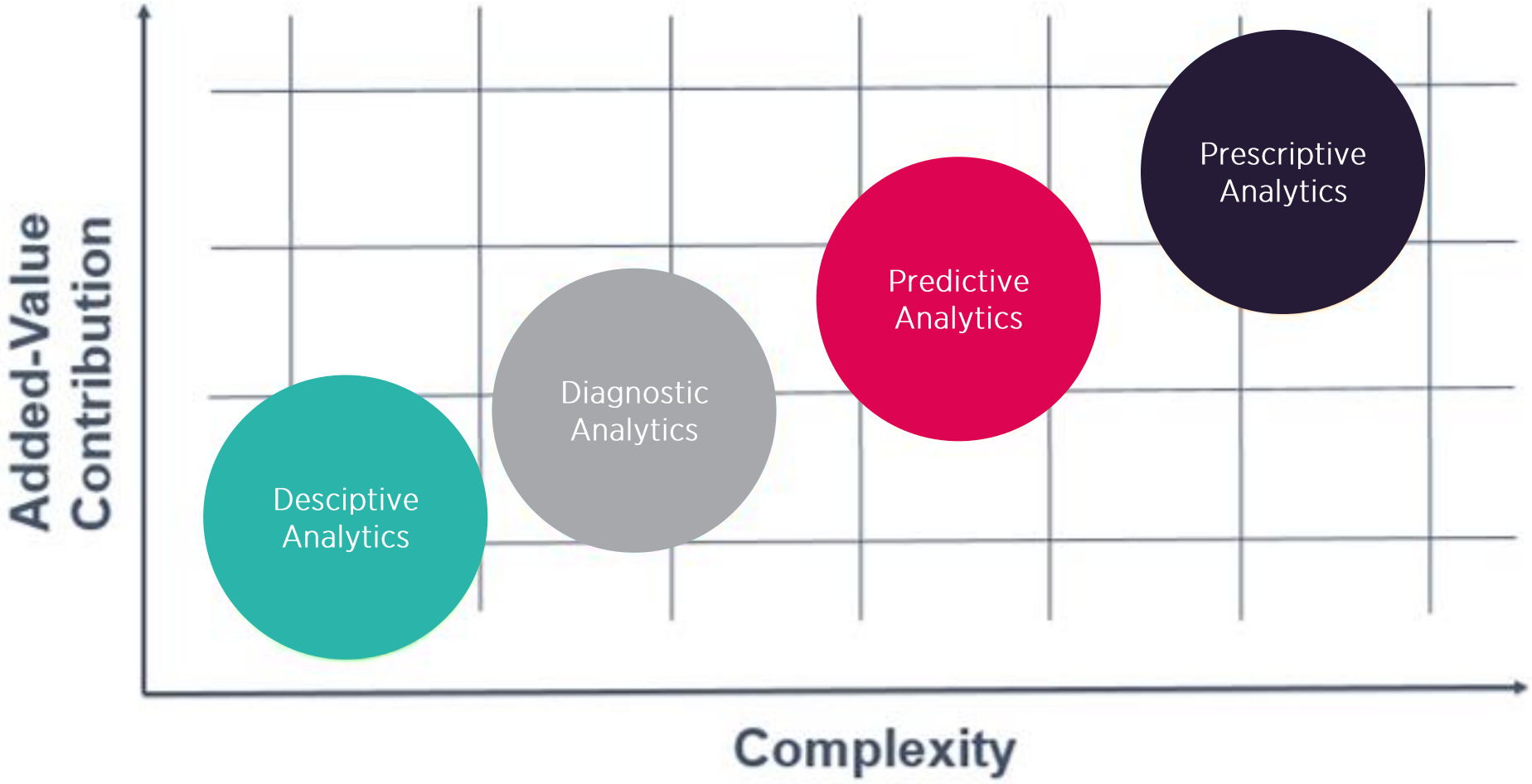
Understand all aspects that make one a complete Data Scientist

- The software programming chops to mess with Data
- The Analytical and Statistical skills to make sense of Data
- All the Machine Learning Theory needed for Data Science
- Real World Understanding of Data and how to understand the business domain to better solve problems



Bütün hüquqlar qorunur.

4 Steps of Data Analytics



Bütün hüquqlar qorunur.

Variables in Data

Data Science Academy

What are the variables?

- Understanding the data we encounter is essential
- Think of data as the information we collect to describe some something

Data can come in two main forms:

- Quantitative variables
- Qualitative or Categorical variables

Bütün hüquqlar qorunur.

Understanding Variable Types

Qualitative /Categorical variable

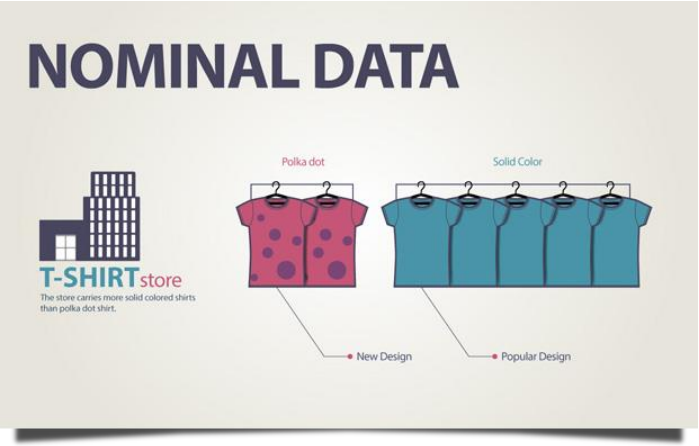
First Name	Last Name	Age	General Subject Area	Overall Grade	Average Mark
Rasul	Jabbarov	18	Sciences	B+	65
Imran	Aliyev	17	Languages	B	62
Namig	Dadashov	17	Modern Arts	C+	56

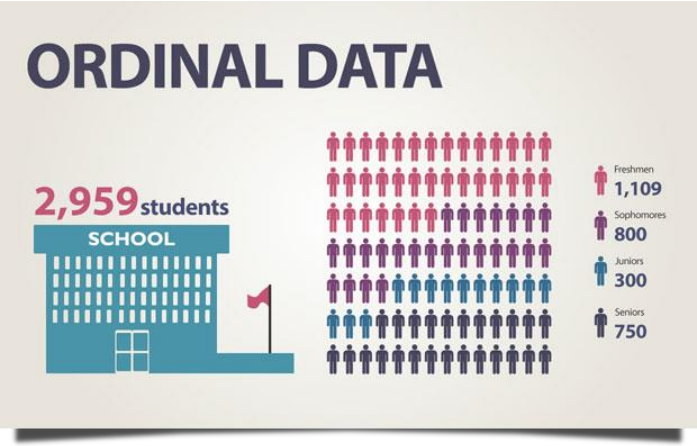
Quantitative variable

Bütün hüquqlar qorunur.

Nominal and Ordinal

- Nominal scale variables differentiate individual data points e.g. an ID variable or Name
- They say nothing about the value, direction, size or any quantitative measure. They are always qualitative.
- Ordinal scale variables can measure direction, size and values. However, e.g. for a high, low, medium measurements. We don't know exact values i.e. how high or how low.





Bütün hüquqlar qorunur.

Interval & Ratio

- **Interval scales** can tell us the size difference between categories, however the still can't tell us the exact value e.g. Date of Birth

- **Ration scales** are similar to interval, but they have the added property of having an inherent zero. E.g. height or weight.

Bütün hüquqlar qorunur.

Continuous and Discrete Variables

- Goals are **discrete**, there is no way a player can score a fraction of a goal. Discrete variables measure quantity and value, but have no interval measurement between adjacent values.

- Height however, is a continuous. Just because we give whole number integer values, that doesn't mean Player 1 and Player 3 are exactly the same height. Player 1 can be 177.3cm while Player 3 can be 176.9cm. You can perhaps never have exactly the same value in height of two people as there would be nanometer differences in height

Player	Goals	Height
Player 1	43	177cm
Player 2	25	180cm
Player 3	3	177cm

Bütün hüquqlar qorunur.

Exploratory Data Analysis

Data Science Academy

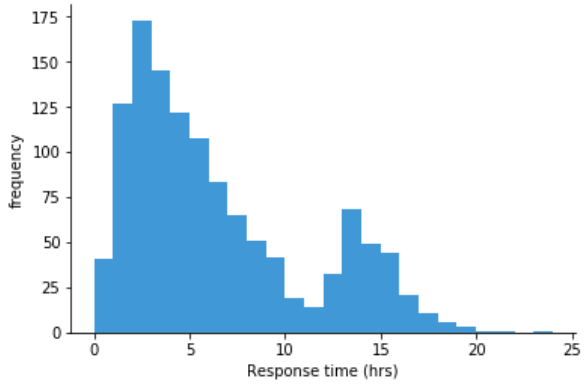
Exploratory Data Analysis (EDA)

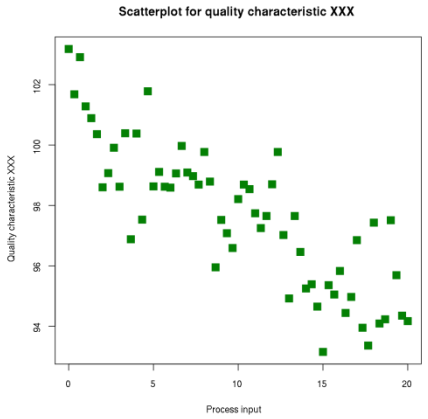
This is the process where we visualize, examine, organize and summarize our data.

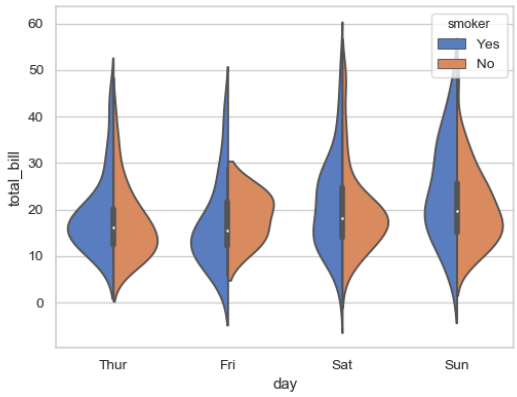
Bütün hüquqlar qorunur.

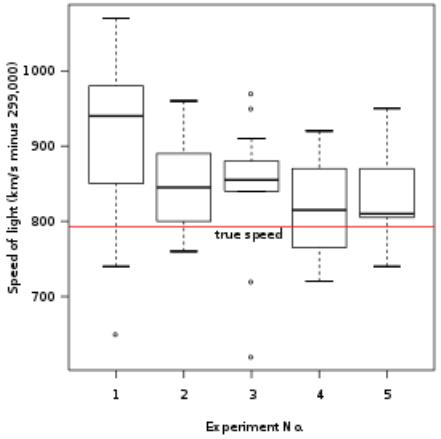
Methods of EDA

- Histogram Plots
- Scatter Plots
- Violin Plots
- Boxplots
- Many more!









Bütün hüquqlar qorunur.

II

Statistics for Data Scientists, Descriptive Statistics

Data Science Academy

Statistics – No one likes statistics

- Almost everyone seems to have a dislike their high school or college Statistics classes
- Most found it confusing, difficult and think it's useless real life.
- They couldn't be more wrong…

Bütün hüquqlar qorunur.

Statistics – Why is so important?

Everyone deals with statistics!

- Will it rain tomorrow?
- Who's expected to score the most goals in the next World Cup?
- Is Trump going to win the next election?

And in business…

- What month will I have the most sales, or what time of day?
- Should I take out insurance?
- Is the economy doing well?

Everything dealing with forecasting/predicting comes down to Statistics

Companies and Researchers leveraging statistical knowledge know:

- What products of movies you like (think Amazon or Netflix)
- When fraudulent activities are taking place
- Predicting customer demand
- Understanding the cause of certain illnesses
- Understanding what the best advertisements, medications, diets and more!
- Everything dealing with forecasting/predicting comes down to Statistics

Bütün hüquqlar qorunur.

“

“Being a Statistician will be the sexiest job over the next decade”

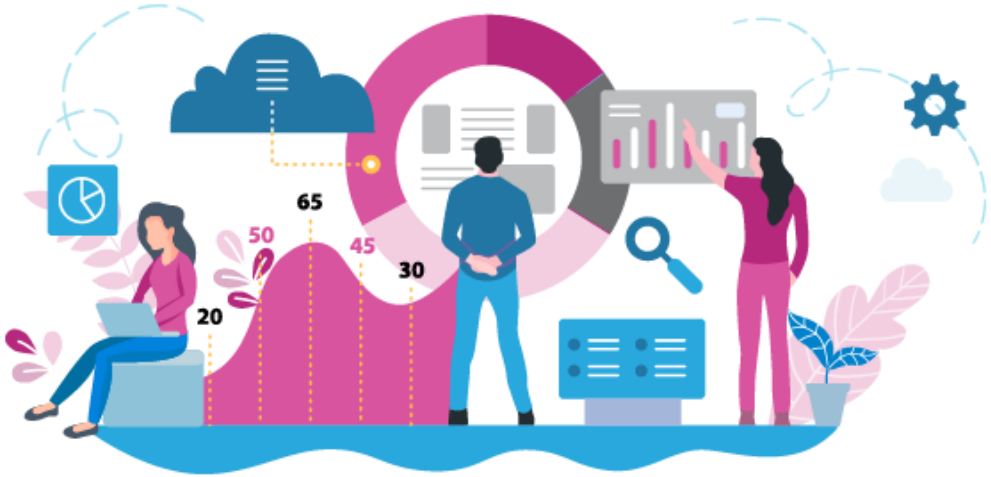
Hal Varian, Chief Economist, Google

Data Science Academy

The Subfields of Statistics

- Descriptive Statistics – Measures or descriptions used to assess some performance or indicator e.g. Batting Averages, GPA

- Inference – Using knowledge from data to make informed inferences, e.g. answering the questions “How often to people get the common cold” or “How many people can afford to a buy a house at the age of 25”



Bütün hüquqlar qorunur.

The Subfields of Statistics

- Risk and Probability – What's the likelihood of your rolling a 6 on a dice? Probability is an extensive and important field that is critical for many businesses such as Insurance, Casinos and Finance.

- Correlation and Relationships – How do we know smoking causes cancer? Extensive statistical studies have to be used for Hypothesis testing. This is an area that's often extremely difficult, but extremely useful in making impactful decisions.

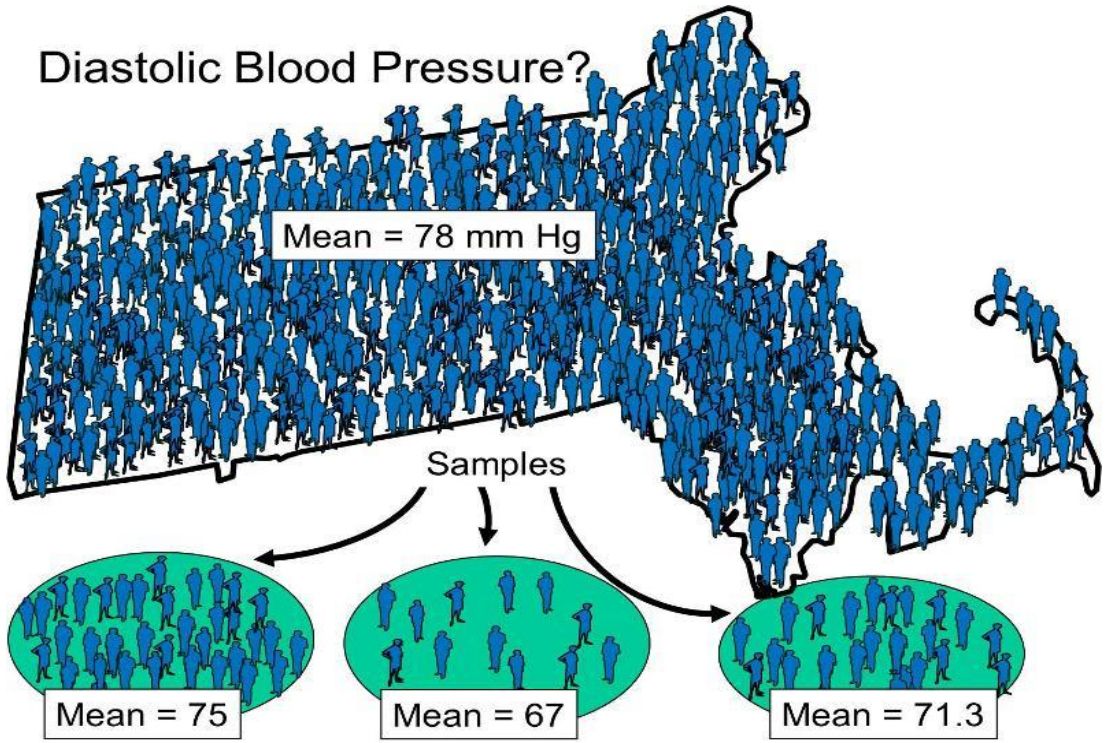
Bütün hüquqlar qorunur.

The Subfields of Statistics

- Modeling – Many times in movies or documentaries, you'll hear scientists referring to "Their model predicts X". In the real world, especially in data science, modeling is the bread and butter of the job. Building good models that predict some outcome based on the inputs, is critical for many industries. E.g. predicting which customers will purchase Item A.

Bütün hüquqlar qorunur.

Basic statistical concepts



Bütün hüquqlar qorunur.

Sampling Summary

- A subset of a **population** (i.e. the total or all individuals in a set) is called a **sample**
- A good sample aims to be a good **representative** form of that population
- **Sampling Error** is the difference between the parameters or descriptive statistics (i.e. mean etc.) of the population. **Small Sampling Errors** are good

Types of Sampling that we use to create good representations include:

- Random sampling
- Satratified sampling

Bütün hüquqlar qorunur.

Descriptive Statistics

Data Science Academy

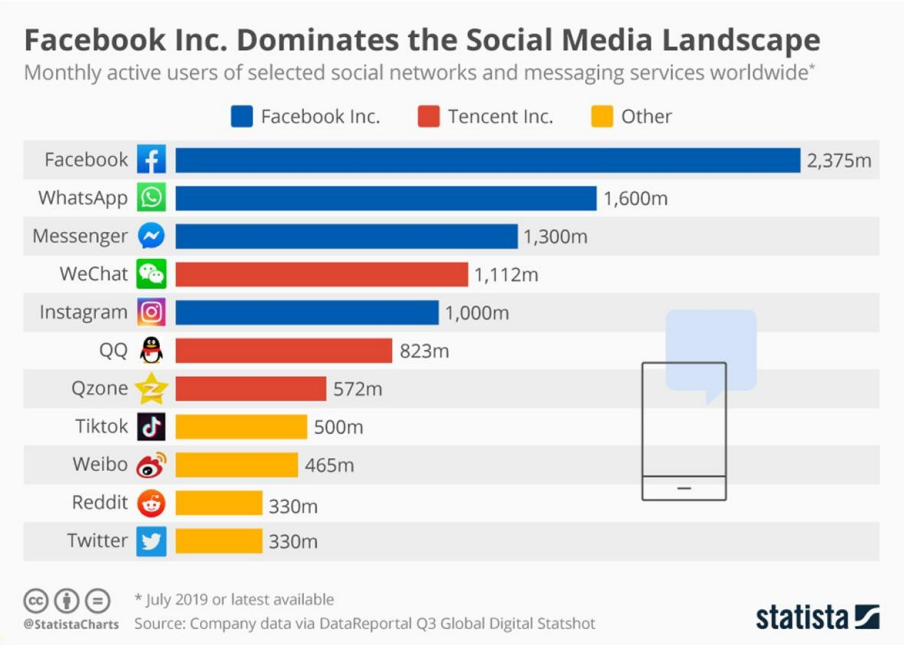
Descriptive Statistics

- Descriptive statistics are used to describe or summarize data in ways that are meaningful and useful, such that, for example, patterns might emerge from the data.

Bütün hüquqlar qorunur.

Examples of descriptive Statistics

- Average heights and weights of males (5’9” and 184lbs for the UK)
- Average number of items sold per day at a store



Bütün hüquqlar qorunur.

Descriptive Statistics: Numerical Methods

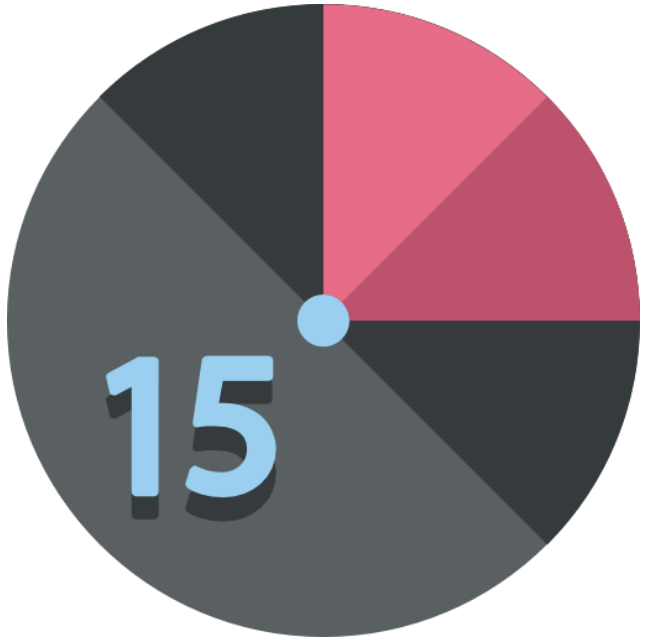
Measures of location

Measures of variability

Bütün hüquqlar qorunur.

Measures of Location

Percentiles

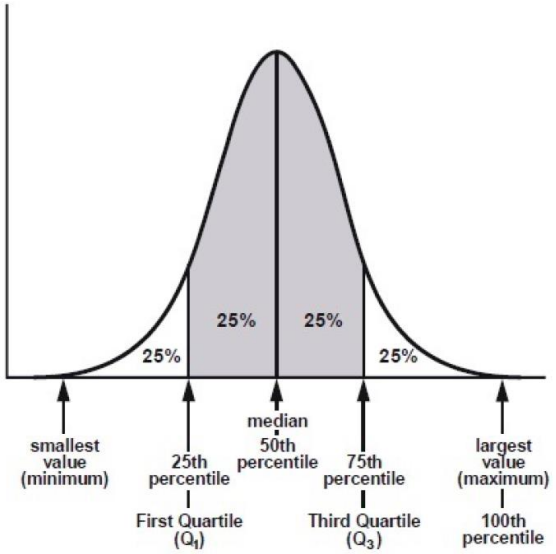


Quartiles

Bütün hüquqlar qorunur.

Percentiles

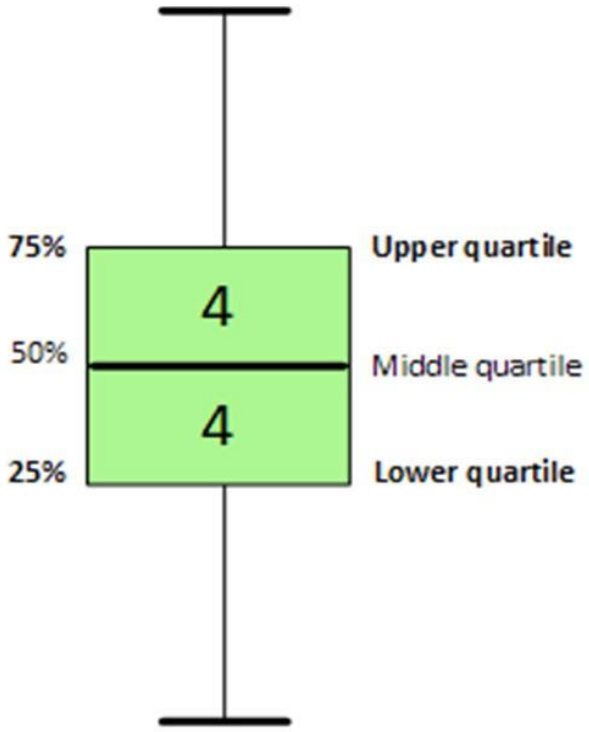
- A percentile provides information about how the data are spread over the interval from the smallest value to the largest value.
- Admission test scores for colleges and universities are frequently reported in terms of percentiles.



Bütün hüquqlar qorunur.

Outliers Rule of Thumb

- A value can be considered an outlier if it exceeds 1.5X the difference between the upper quartile and the lower quartile (the inter Quartile range).



Inter Quartile Range = Upper quartile - Lower quartile

Bütün hüquqlar qorunur.

Measures of location

Mean

The mean is the most widely spread measure of central tendency. It is the simple average of the dataset. (Easily affected by outliers)
The formula to calculate the mean is:

- $\frac{{}^{N}_{i=0} Xi}{N}$ or $\frac{X1+X2+X3+\cdots+XN-1+XN}{N}$

In Excel, the mean is calculated by:
=AVERAGE()

Median

The median is the midpoint of the ordered dataset. It is not as popular as the mean, but is often used in academia and data science. That is since it is not affected by outliers.
In an ordered dataset, the median

is the number at position $\frac{n+1}{2}$

If this position is not a whole

number, it, the median is the simple average of the two numbers at positions closest to the calculated value.

In Excel, the mean is calculated by:
=MEDIAN()

Mode

The mode is the value that occurs most often. A dataset can have 0 modes, 1 mode or multiple modes. The mode is calculated simply by finding the value with the highest frequency.

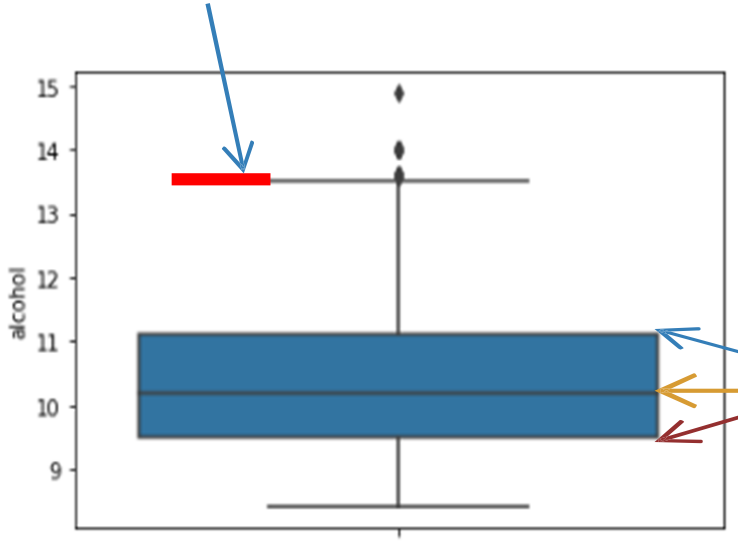
In Excel, the mean is calculated by:
=MODE.SNGL() -> returns one mode

=MODE.MULT() -> returns an array with the modes. It is used when we have more than 1 mode.

Bütün hüquqlar qorunur.

Inter Quartile Ranges for

Upper Outlier Range



Values exceeding 13.5 or being less than 7.1 are outliers

Red Wines

```
# Showing the quartile ranges of alcohol content

df[df['type'] == 'red']['alcohol'].describe()
```

```
count    1599.000000
mean       10.422983
std         1.065668
min         8.400000
25%         9.500000
50%        10.200000
75%        11.100000
max        14.900000
Name: alcohol, dtype: float64
```

Upper Quartile = 11.1
Lower Quartile = 9.5

Inter Quartile Range = 11.1 – 9.5 = 1.6

Outlier Ranges = 1.6 * 1.5 = 2.4

Upper Outlier Range = 11.1 + 2.4 = 13.5

Lower Outlier Range = 9.5 - 2.4 = 7.1

Bütün hüquqlar qorunur.

The Mean

- The mean is quite simple and we will have discussed it before so you intuitively know that mean is the same as average.

$$\text{Mean} = \frac{10 + 16 + 7}{3} = 11$$

Person	ARPU
Naila	10
Afag	16
Sanan	7
Mean	**11**

Bütün hüquqlar qorunur.

Median

- Many people confuse Means and Medians.
- Remember while means are the average of all values, Median is average of the two middle values or the actual middle value itself (depending on if the quantity of data is odd or even)

4	6	7	**8**	10	11	30
			MEDIAN			

Mean= 10.86

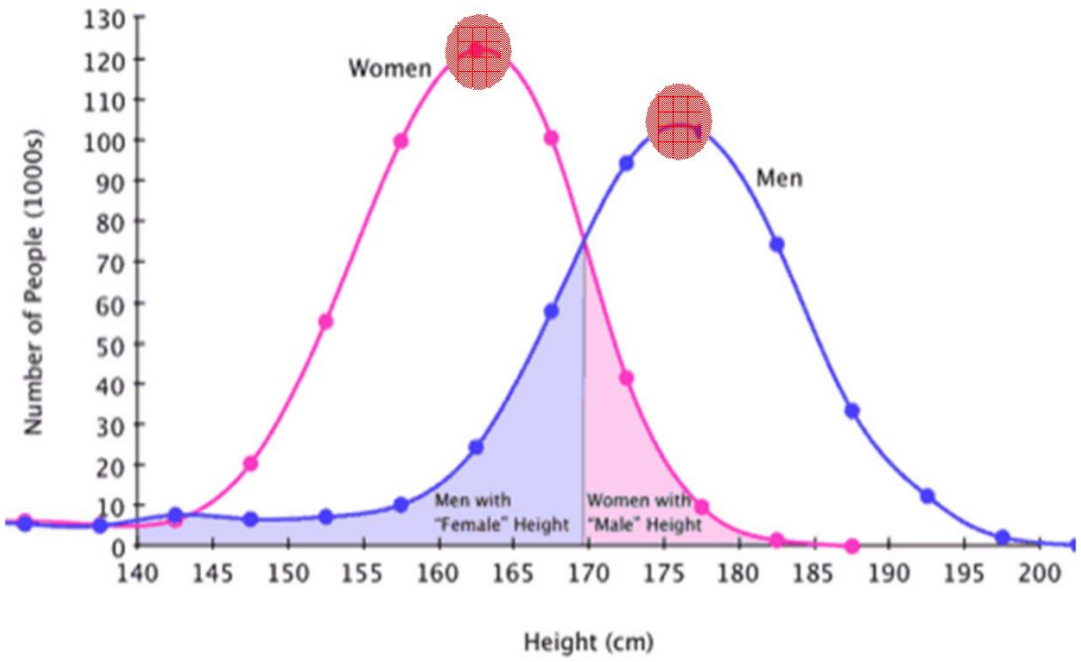
4	6	7	**8**	**10**	11	14	41
			MEDIAN = (8+10)/2 = 9				

Mean= 12.63

Bütün hüquqlar qorunur.

Mode

- The Mode is simply the most frequent item in distribution.
- In any Kernel Density plot (KDE in Seaborn), the mode is always the peak



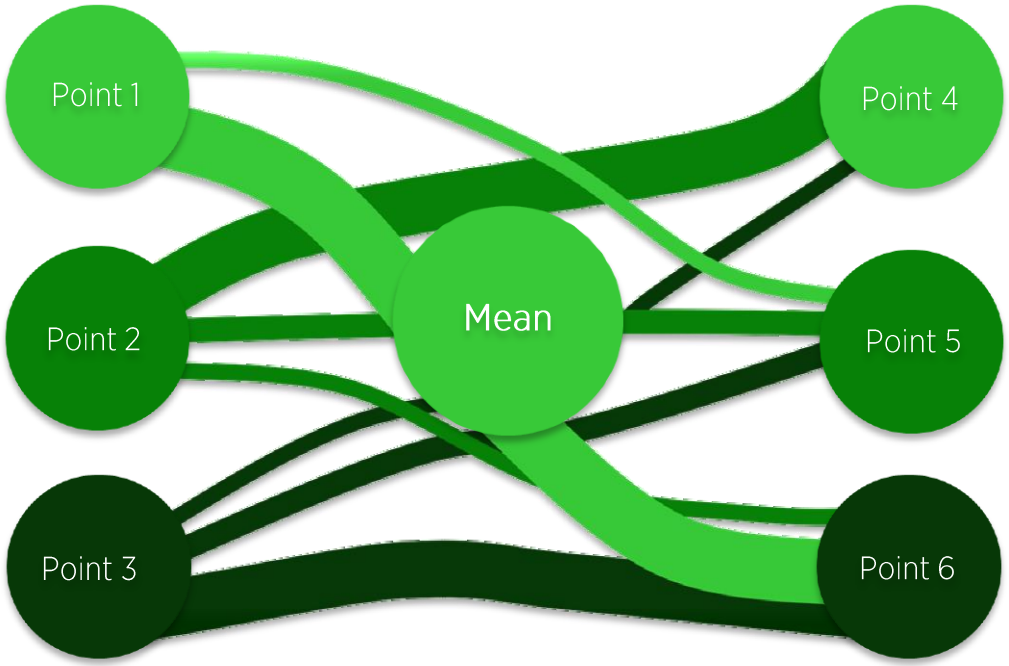
Bütün hüquqlar qorunur.

Measures of Variability

- It is often desirable to consider measures of variability (dispersion), as well as measures of location.
- For example, in choosing supplier A or supplier B, we might consider not only the average delivery time for each, but also the variability in delivery time for each.

Bütün hüquqlar qorunur.

Measures of Variability



Calculating variance in Excel:

Simple variance: =VAR.S()

Population variance: =VAR.P()

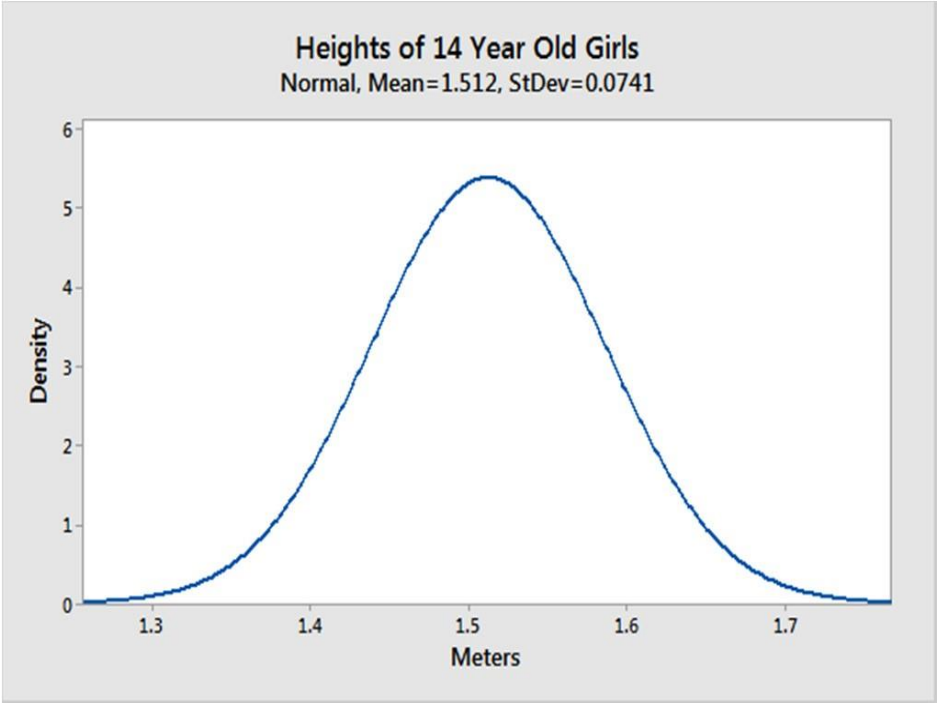
Simple standard deviation: =STDEV.S()

Population standard deviation: =STDEV.P()

- Variance and standard deviation measure the dispersion of a set of data points around ist mean value.
- There are different formulas for population and sample variance & standard deviation. This is due to the fact that the sample formulas are the unbiased estimators of the population formulas.
- Sample variance formula $s^2 = \frac{\sum_{i=1}^{n} (x_i - x)^2}{n-1}$
- Population variance formula $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$
- Sample standard deviation formula $s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - x)^2}{n-1}}$
- Population standard deviation formula $\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$

Bütün hüquqlar qorunur.

Normal Distributions Example



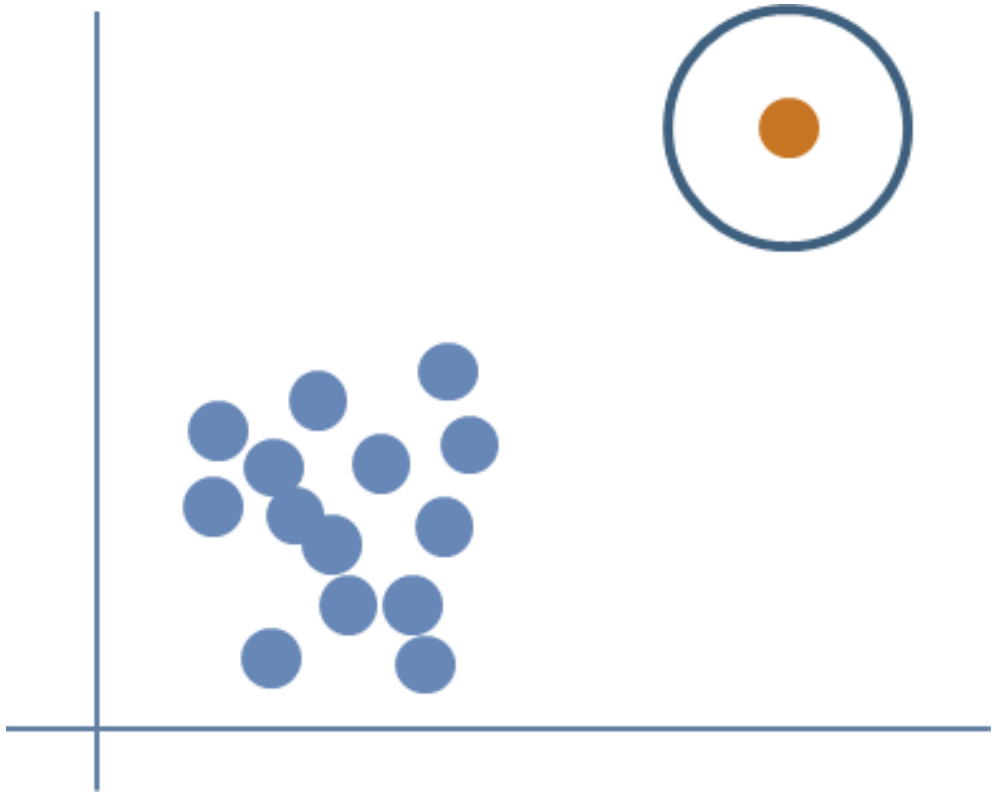
Data Science Academy

Detecting Outliers

Data Science Academy

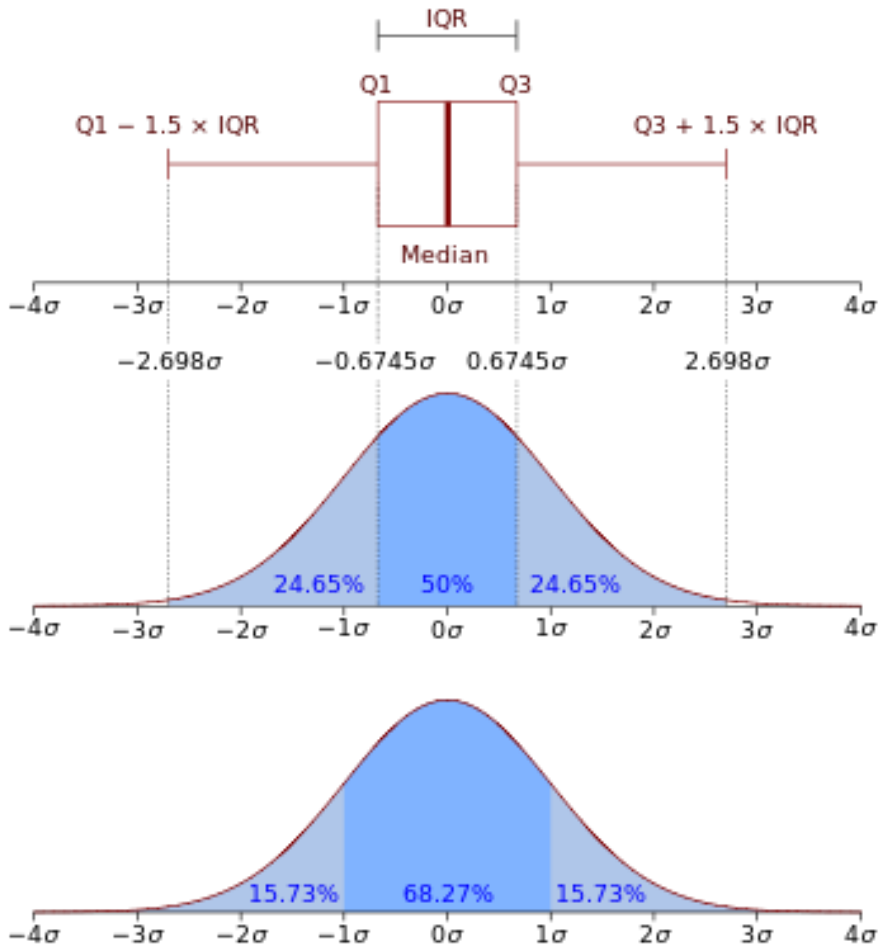
Detecting Outliers

- An outlier is an unusually small or unusually large value in a data set.
- A data value with a z-score less than -3 or greater than +3 might be considered an outlier.
- It might be an incorrectly recorded data value.
- It might be a data value that was incorrectly included in the data set.
- It might be a correctly recorded data value that belongs in the data set !



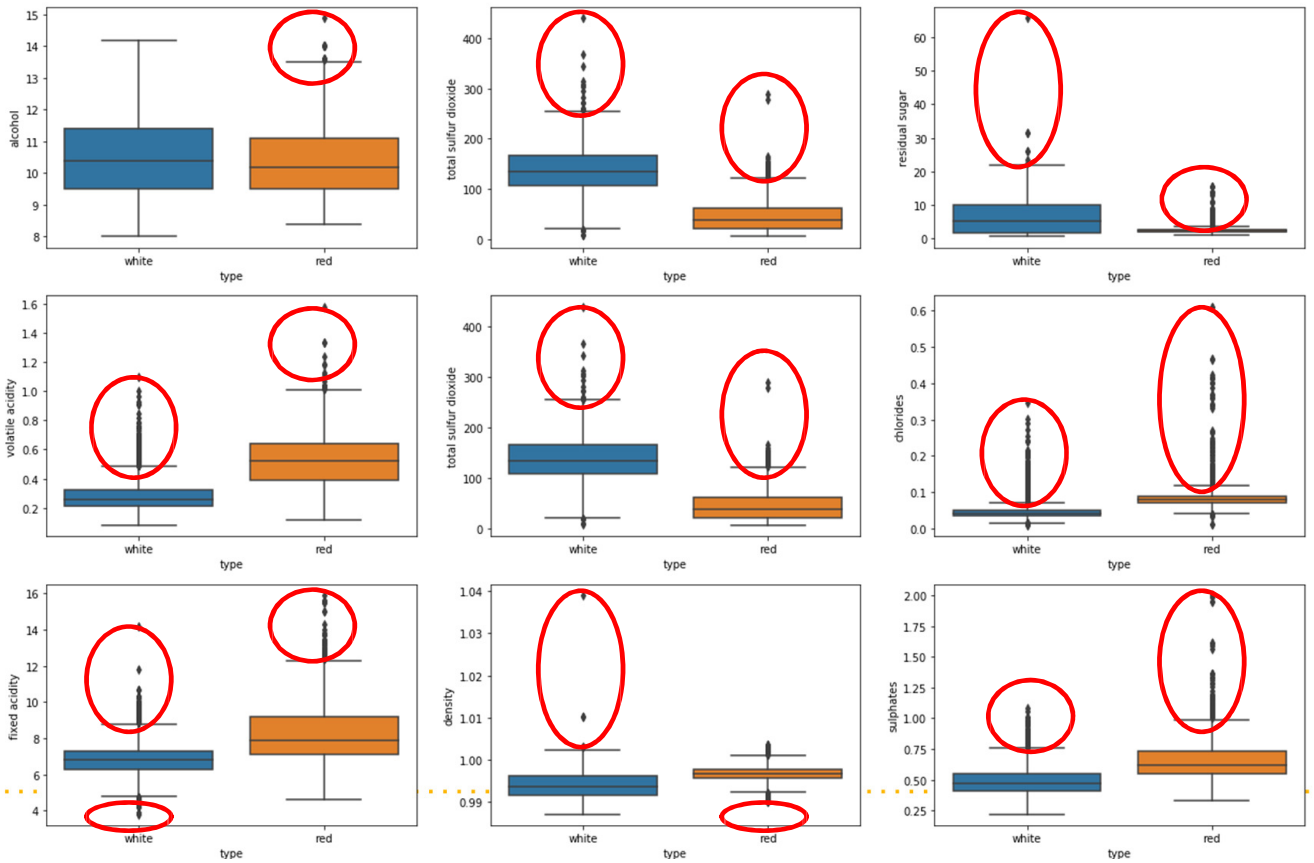
Bütün hüquqlar qorunur.

Detecting Outliers



Bütün hüquqlar qorunur.

Analyze Frequency Distributions to Find Outliers



Bütün hüquqlar qorunur.

III

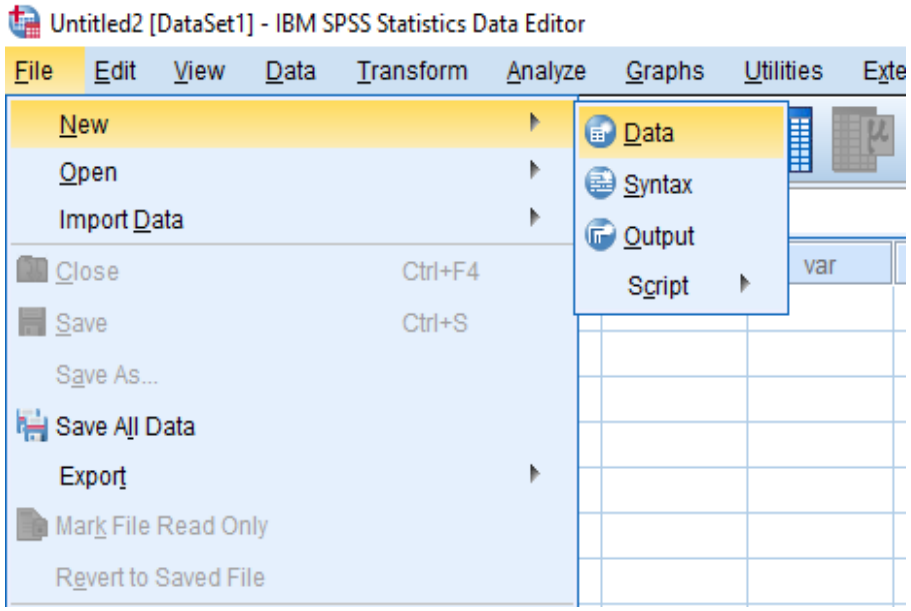
Introduction to SPSS Statistics and SPSS Modeler

Data Science Academy

Creating and opening files in SPSS

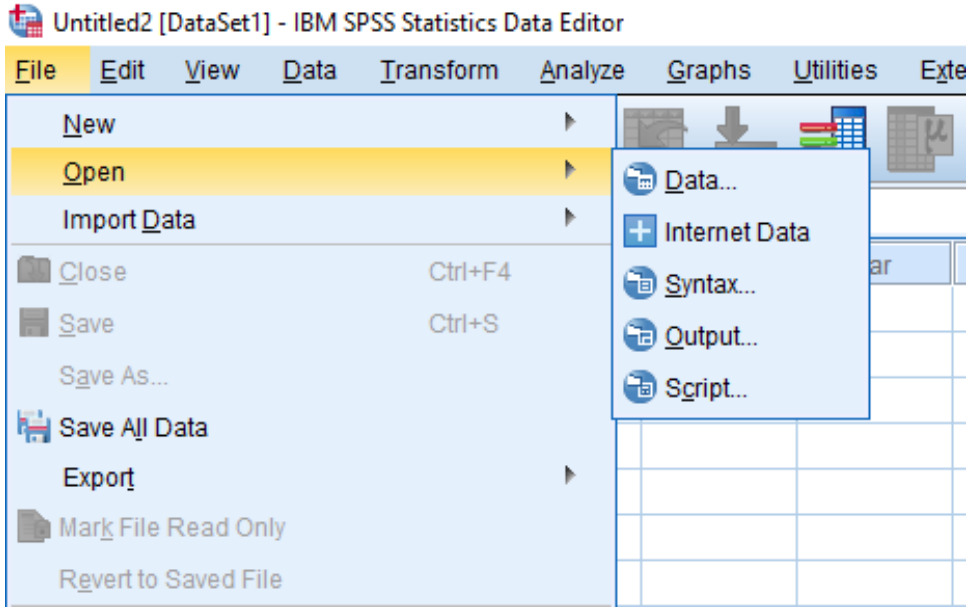
File → New → Data

1.



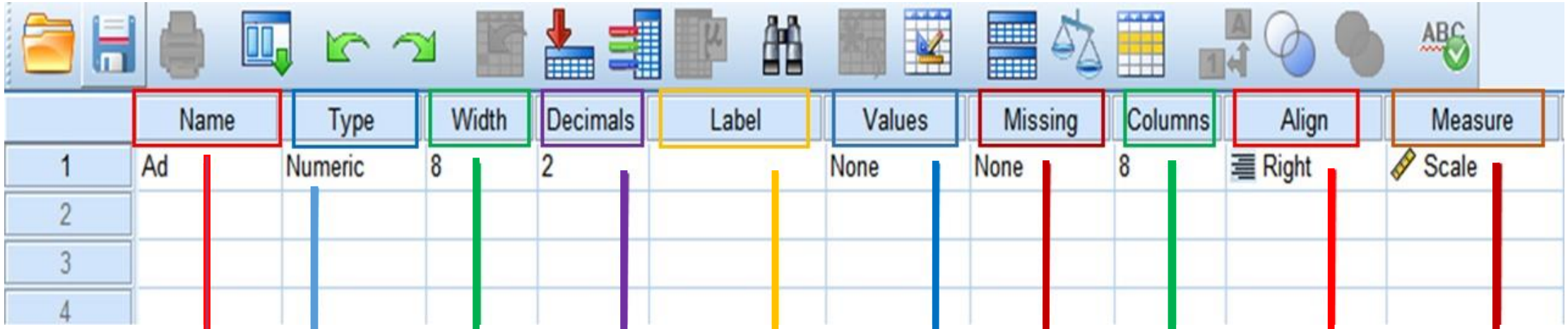
File → Open → Data

2.



Bütün hüquqlar qorunur.

Adding data in the SPSS



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	Ad	Numeric	8	2		None	None	8	Right	Scale
2										
3										
4										

Under the Name, we establish the title (header) of the variable. Example: Age, Surname, Job Title.

Under the Type, we select the type of the variable. Example: For text and characters we choose Text, for numeric variables – Numeric.

Width shows us the maximum number of characters and digits for particular variable.

Decimals shows us the number of decimal places in fractional number.

Label shows additional explanation of the variable.

Under the Values you can give a text label for category codes by clicking on the appropriate cell in the column values. Default is None.

Missing is used to define different type of missing value code.

The width of column in the data view sheet can be altered by clicking in the appropriate Column cell.

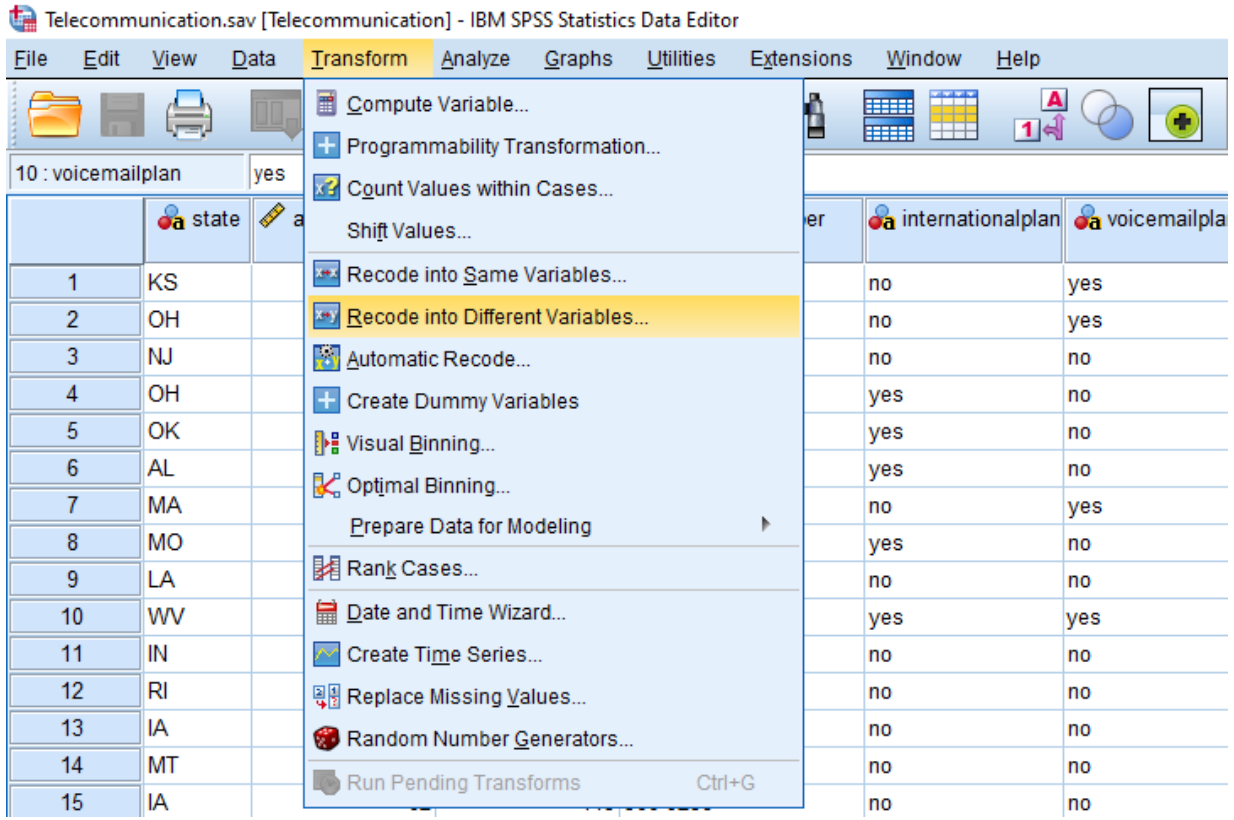
To align the text or numbers in Data View, click in the appropriate Align cell.

To label the scale of data, click in the appropriate Measure cell, “Ordinal”, “Nominal” or in our example “Scale”.

Bütün hüquqlar qorunur.

Practice: Recoding of the variables (Telecommunication data)

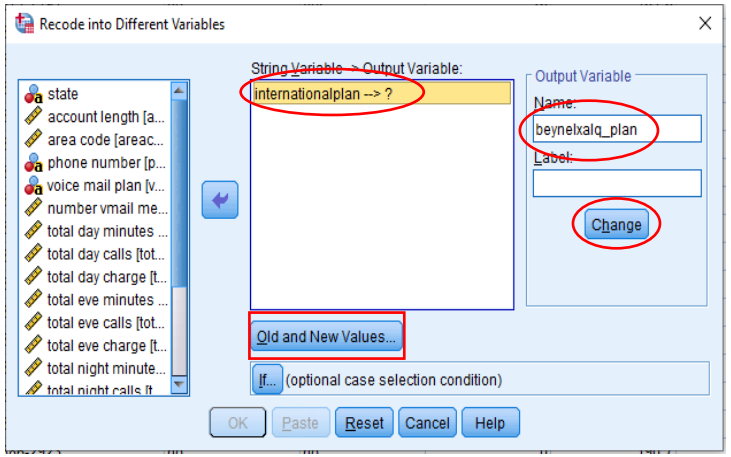
Transform → Recode into Different Variables



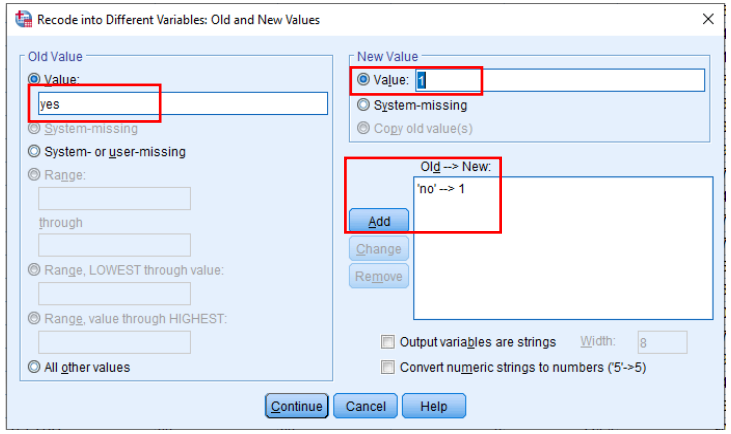
Bütün hüquqlar qorunur.

Practice: Recoding of the variables (Telecommunication data)

1)



2)



- Drag **internationalplan** variable into box*
- Enter an output (new) variable name for each new variable and click **Change**
- Click **Old and New Values**
- Specify an old value and a new value.
- Click **Add** to place the specification into the Old–>New list
- Click **«Continue»**
- Click **«OK»**

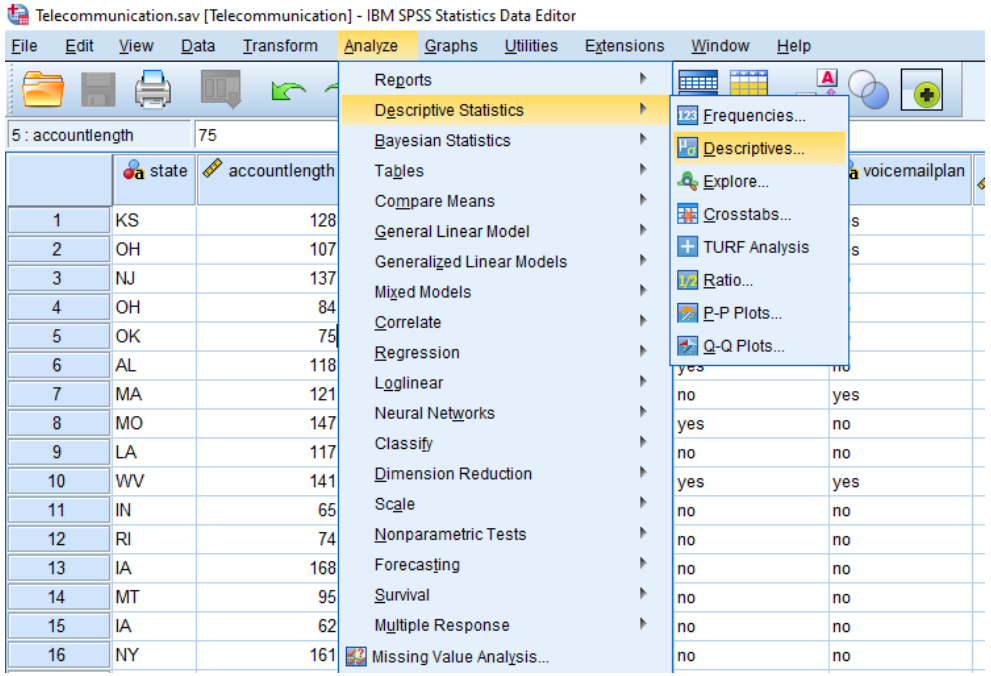
  - If you select multiple variables, they must be the same
  - type (numeric or string).

Bütün hüquqlar qorunur.

Practice: Detecting Outliers – Numerical Method (Telecommunication data)

Task: check if *total_day_minutes* has outliers

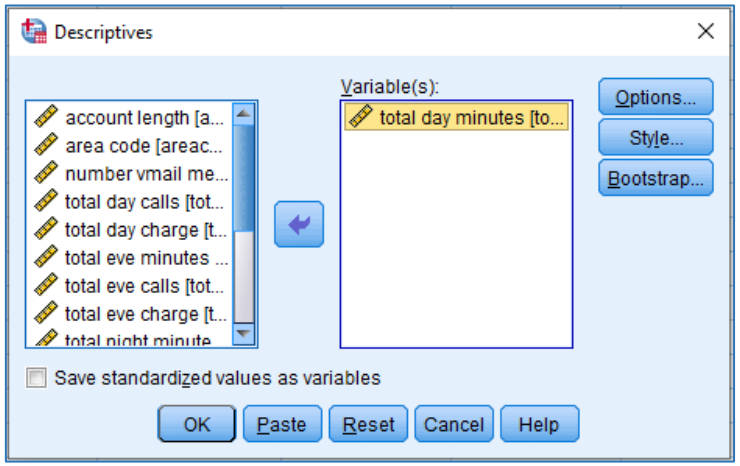
Analyze → Descriptive Statistics → Descriptives



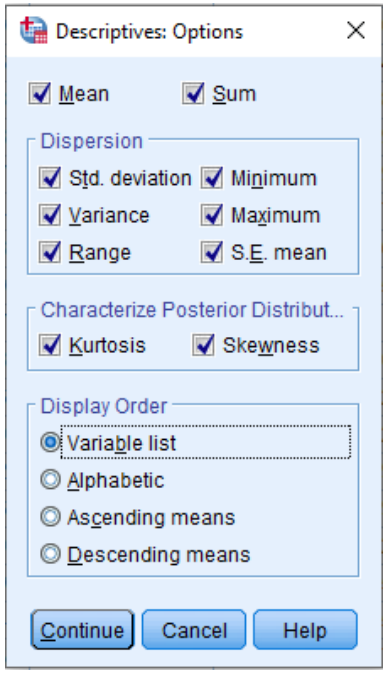
Bütün hüquqlar qorunur.

Describe total_day_minutes variable using SPSS Statistics

1)



2)

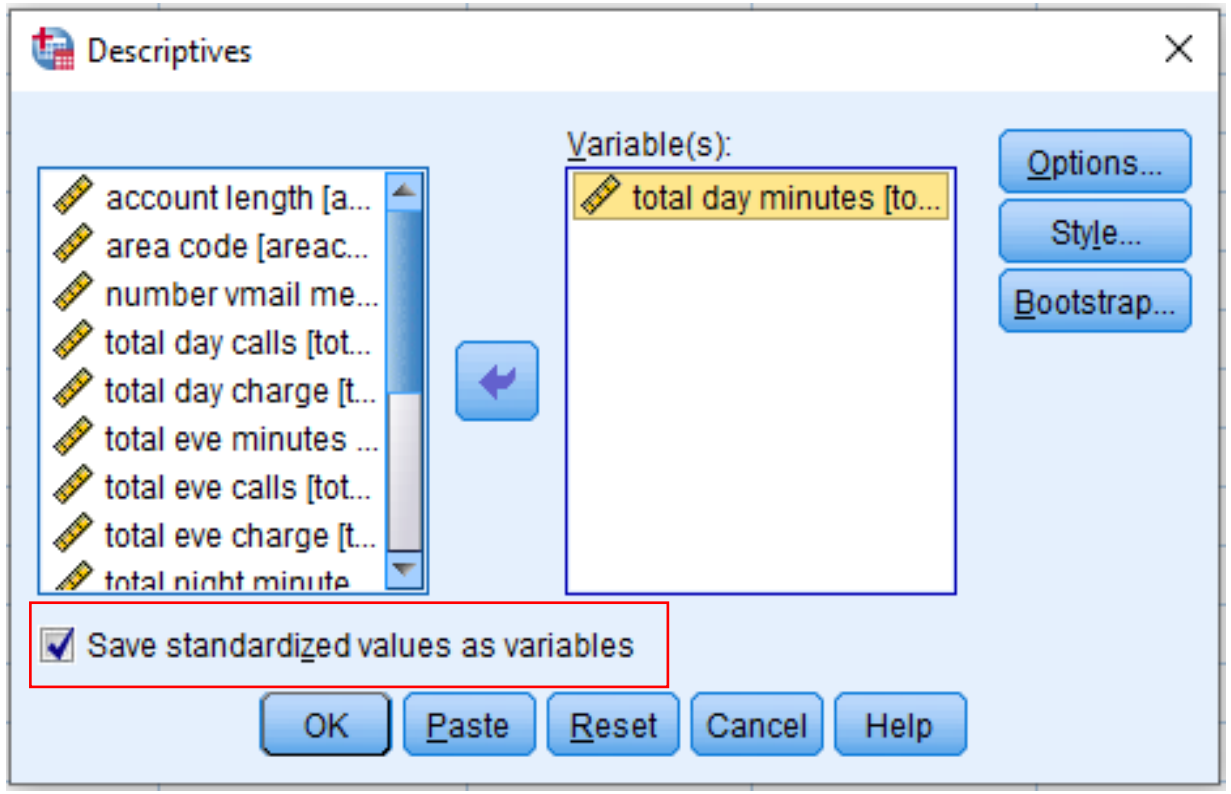


- Drag **total day minutes** variable into box*
- Click **Options**
- Tick Mean, Sum, Std.deviation, Minimum, Maximum, Variance, Range, S.E.mean, Kurtosis, Skewness
- Click **«Continue»**
- Click **«OK»**

* If you select multiple variables, they must be the same type (numeric or string).

Bütün hüquqlar qorunur.

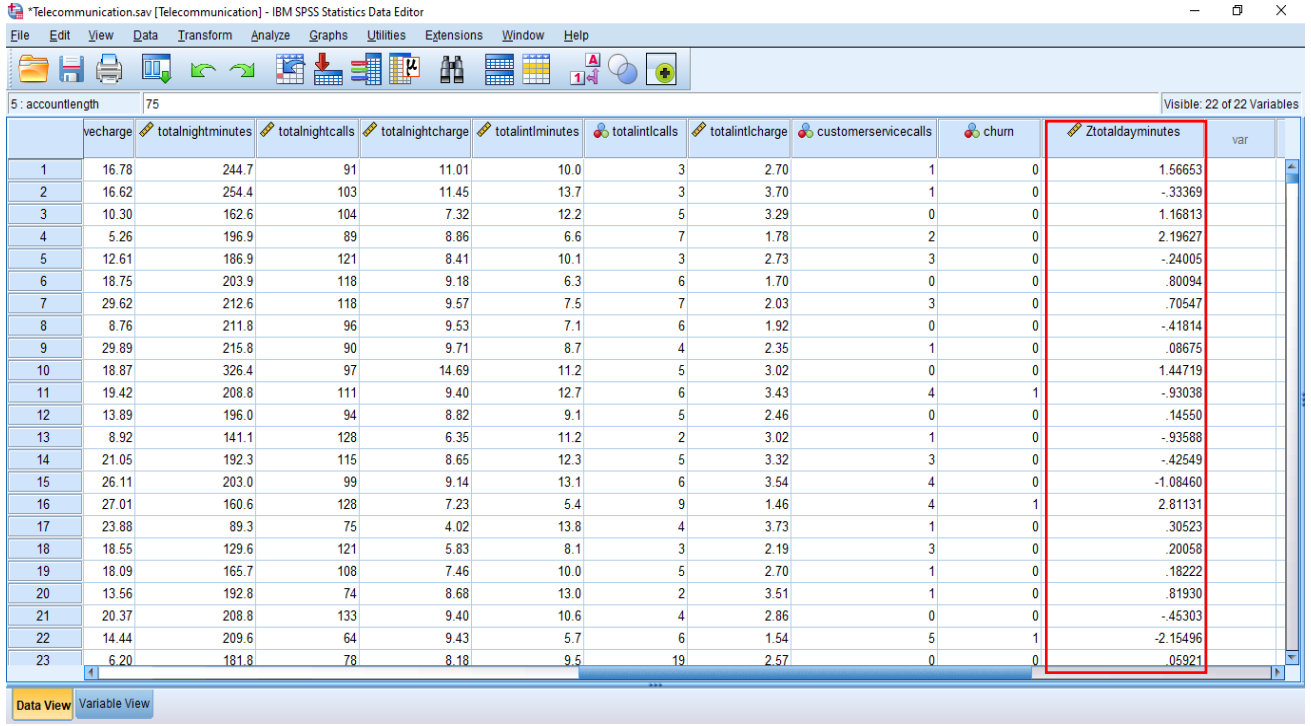
Z-Scores



- Drag **total day minutes** variable into box
- Tick **Save Standardized values as variables**
- Select **«Ok»**

Bütün hüquqlar qorunur.

Detecting outliers- Z-scores



	vecharge	totalnightminutes	totalnightcalls	totalnightcharge	totalintlminutes	totalintlcalls	totalintlcharge	customerservicecalls	churn	Ztotaldayminutes	var
1	16.78	244.7	91	11.01	10.0	3	2.70	1	0	1.56653	
2	16.62	254.4	103	11.45	13.7	3	3.70	1	0	-.33369	
3	10.30	162.6	104	7.32	12.2	5	3.29	0	0	1.16813	
4	5.26	196.9	89	8.86	6.6	7	1.78	2	0	2.19627	
5	12.61	186.9	121	8.41	10.1	3	2.73	3	0	-.24005	
6	18.75	203.9	118	9.18	6.3	6	1.70	0	0	.80094	
7	29.62	212.6	118	9.57	7.5	7	2.03	3	0	.70547	
8	8.76	211.8	96	9.53	7.1	6	1.92	0	0	-.41814	
9	29.89	215.8	90	9.71	8.7	4	2.35	1	0	.08675	
10	18.87	326.4	97	14.69	11.2	5	3.02	0	0	1.44719	
11	19.42	208.8	111	9.40	12.7	6	3.43	4	1	-.93038	
12	13.89	196.0	94	8.82	9.1	5	2.46	0	0	.14550	
13	8.92	141.1	128	6.35	11.2	2	3.02	1	0	-.93588	
14	21.05	192.3	115	8.65	12.3	5	3.32	3	0	-.42549	
15	26.11	203.0	99	9.14	13.1	6	3.54	4	0	-1.08460	
16	27.01	160.6	128	7.23	5.4	9	1.46	4	1	2.81131	
17	23.88	89.3	75	4.02	13.8	4	3.73	1	0	.30523	
18	18.55	129.6	121	5.83	8.1	3	2.19	3	0	.20058	
19	18.09	165.7	108	7.46	10.0	5	2.70	1	0	.18222	
20	13.56	192.8	74	8.68	13.0	2	3.51	1	0	.81930	
21	20.37	208.8	133	9.40	10.6	4	2.86	0	0	-.45303	
22	14.44	209.6	64	9.43	5.7	6	1.54	5	1	-2.15496	
23	6.20	181.8	78	8.18	9.5	19	2.57	0	0	.05921	

- Go back to **Data View**
- Right click on **Ztotaldayminutes**
- Select descending

Bütün hüquqlar qorunur.

Introduction to SPSS Modeler

Data Science Academy

SPSS Modeler

SPSS Modeler is a leading visual data science and machine-learning solution. It helps enterprises accelerate time to value and achieve desired outcomes by speeding up operational tasks for data scientists.

IBM
SPSS MODELER
PROFESSIONAL

Bütün hüquqlar qorunur.

SPSS Modeler

SPSS Modeler helps you:

- Take advantage of open source-based innovation, including R or Python.
- Empower data scientists of all skills — programmatic and visual.
- Explore a hybrid approach — on premises and in the public or private cloud.
- Start small and scale to an enterprise-wide, governed approach.

Bütün hüquqlar qorunur.

Business benefits of SPSS Modeler

Go live in days, not months or weeks

Start with GUI-based data science and machine learning algorithms.

Scale with peace of mind

Start small and expand with enterprise-class security and governance.

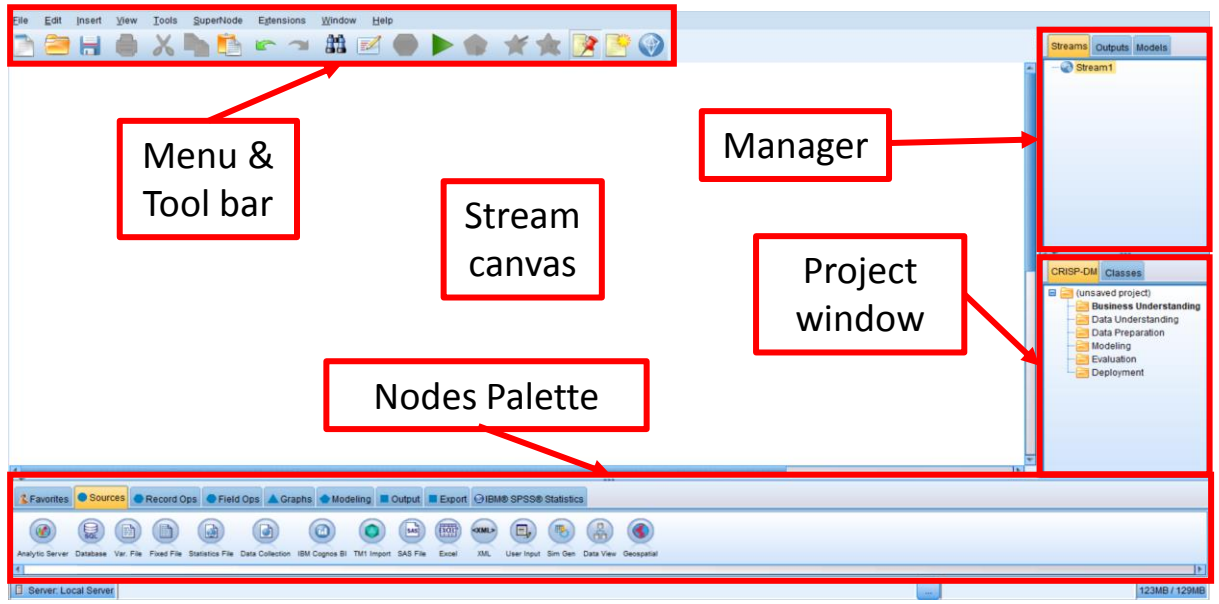
Boost data science team productivity

Empower coders, non-coders and analysts.

Bütün hüquqlar qorunur.

SPSS Modeler Interface

- Stream canvas - is where you build and manipulate data streams.
- Managers - has three tabs, which are used to manage streams, output and models. You can use Streams tab to open, rename, save, and delete created streams in a session.
- Project window - used to create and manage data mining projects (groups of files related to a data mining task).



Bütün hüquqlar qorunur.

Toolbar Icons

Create new stream

Copy to clipboard

Open stream

Redo

Save stream

Run current stream

Run stream selection

Add SuperNode

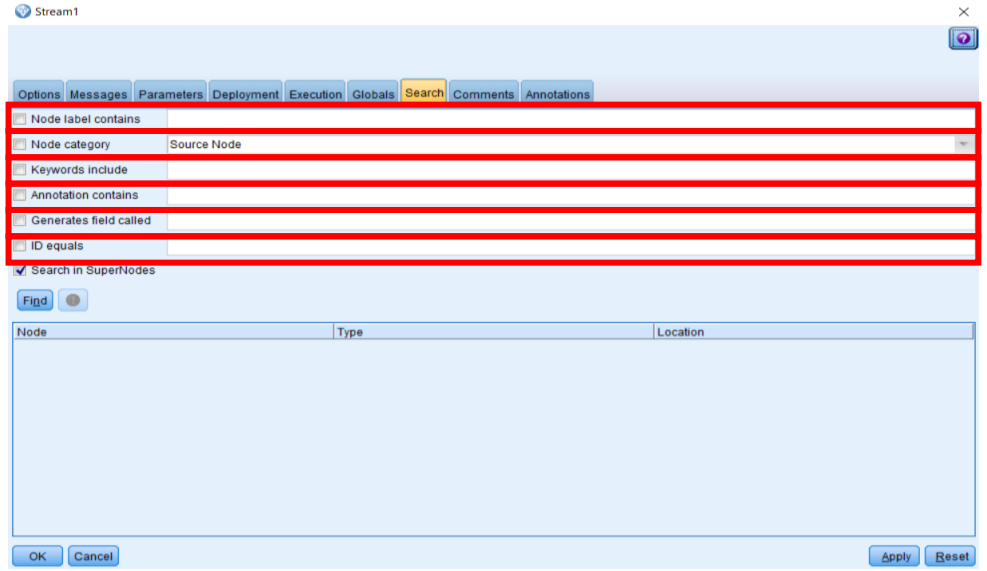
Stop stream (Active only while stream is running)

Search for nodes

Bütün hüquqlar qorunur.

Search for Nodes

- **Node label contains.** Check this box and enter all or part of a node label to search for a particular node.
- **Node category.** Check this box and click a category on the list to search for a particular type of node.
- **Keywords include.** Check this box and enter one or more complete keywords to search for nodes having that text in the Keywords field on the Annotations tab of the node dialog box.
- **Annotation contains.** Check this box and enter one or more words to search for nodes that contain this text in the main text area on the Annotations tab of the node dialog box.
- **Generates field called.** Check this box and enter the name of a generated field (for example, $C-Drug).
- **ID equals.** Check this box and enter a node ID to search for a particular node with that identifier (selecting this option disables all the preceding options).



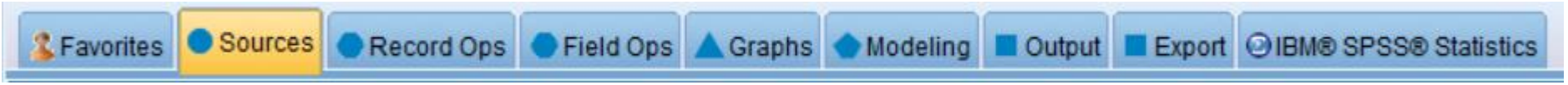
Bütün hüquqlar qorunur.

Nodes Palette

Nodes Palette consists of :

- Sources
- Record ops
- Field ops
- Graphs
- Modelling
- Output
- Export

- Data Manipulations nodes
- Output nodes



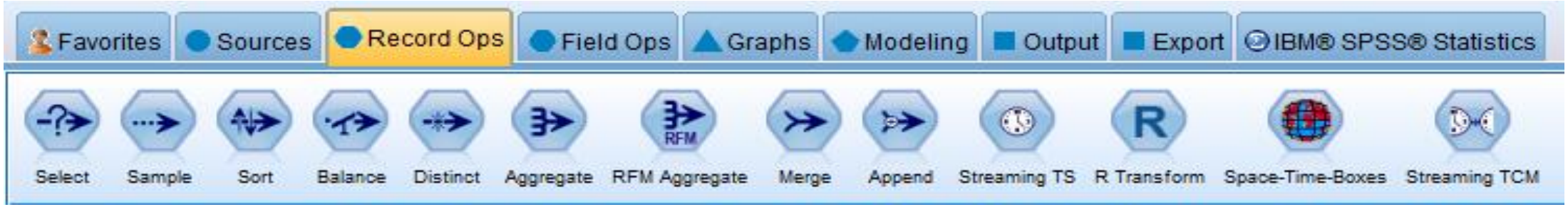
Bütün hüquqlar qorunur.

Nodes and Supernodes

Data Science Academy

What is Node?

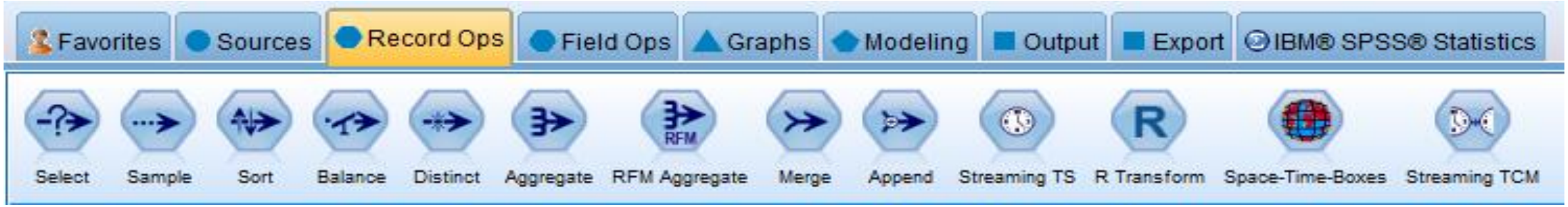
Nodes are used in IBM SPSS Modeler to help you explore data. Various nodes in the workspace represent different objects and actions.



Bütün hüquqlar qorunur.

What is Node?

Nodes are used in IBM SPSS Modeler to help you explore data. Various nodes in the workspace represent different objects and actions.



Bütün hüquqlar qorunur.

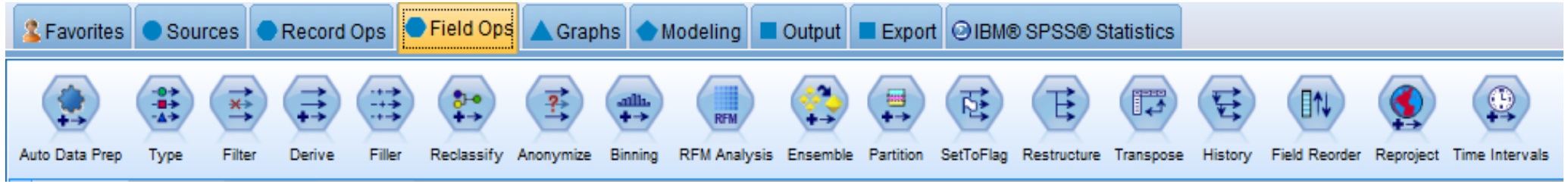
Source Nodes

Source nodes bring data into the stream, and are located on the Sources tab of the nodes palette

Bütün hüquqlar qorunur.

Data Manipulations

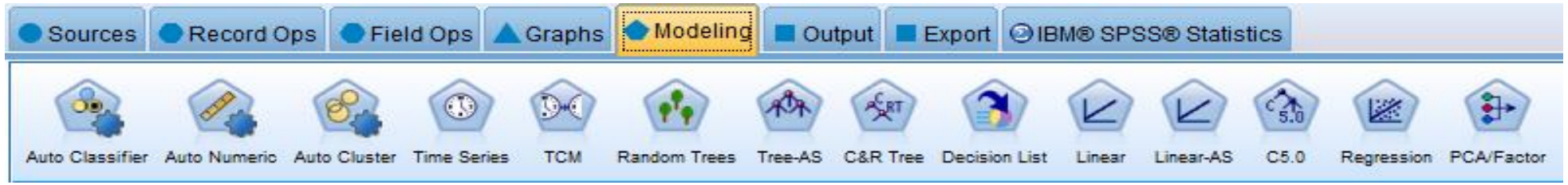
Process nodes perform operations on individual data records and fields, and can be found in the Record Ops and Field Ops tabs of the palette.



Bütün hüquqlar qorunur.

Data Manipulations

Modeling nodes use statistical algorithms to create model nuggets, and are located on the Modeling tab, and (if activated) the Database Modeling tab, of the nodes palette.



Bütün hüquqlar qorunur.

Visualize Data

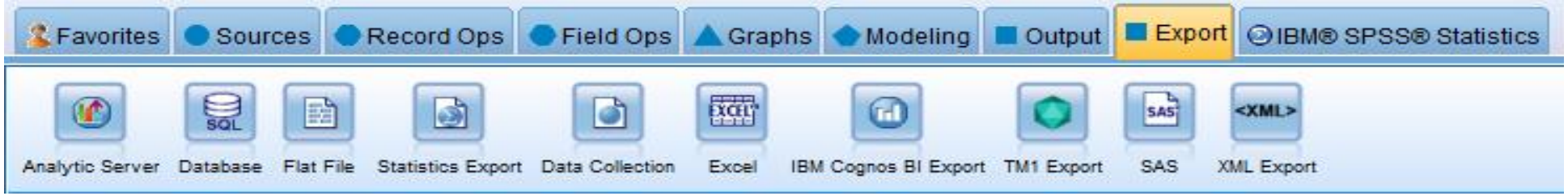
Graph nodes graphically display data before and after modeling. Graphs include plots, histograms, web nodes, and evaluation charts.

Bütün hüquqlar qorunur.

Output

Output nodes produce a variety of output for data, charts and model results, and are included on the Graphs, Output and Export tabs of the nodes palette.

Export nodes produce a variety of output that can be viewed in external applications, such as IBM® SPSS® Data Collection or Excel.

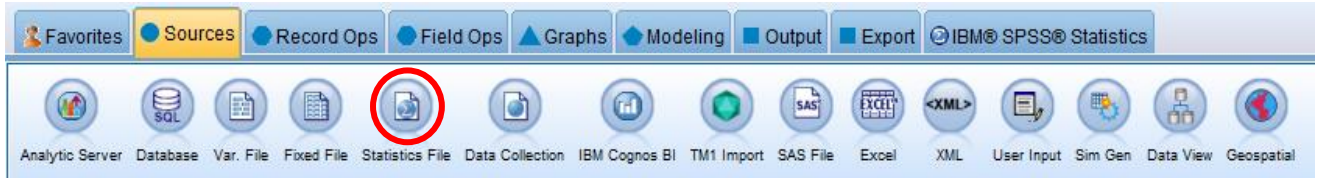


Bütün hüquqlar qorunur.

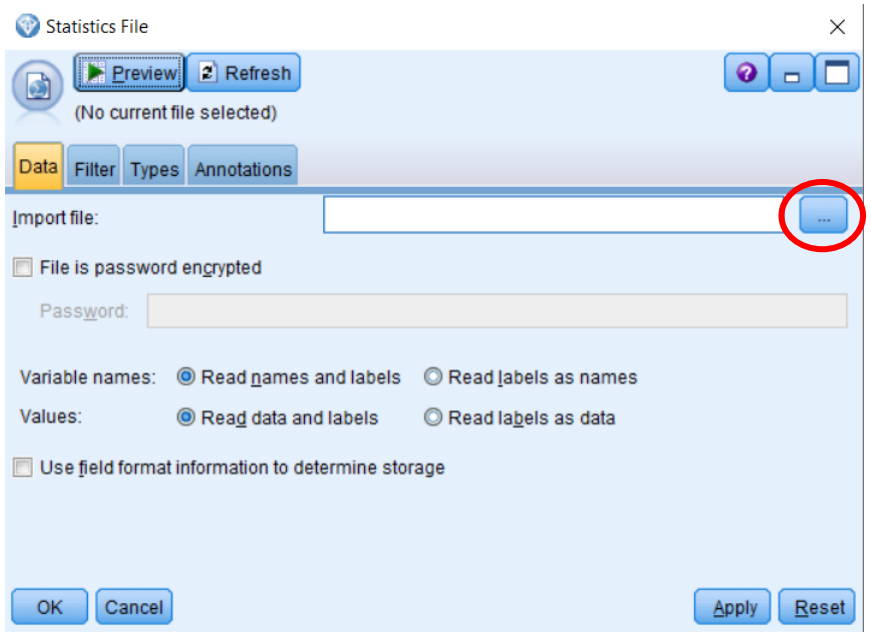
Connecting Data

1. Double click on the **Statistics File** from **Sources** tab;
2. Double click on the **Statistics File** on the canvas;
3. Click ... button;

1



2

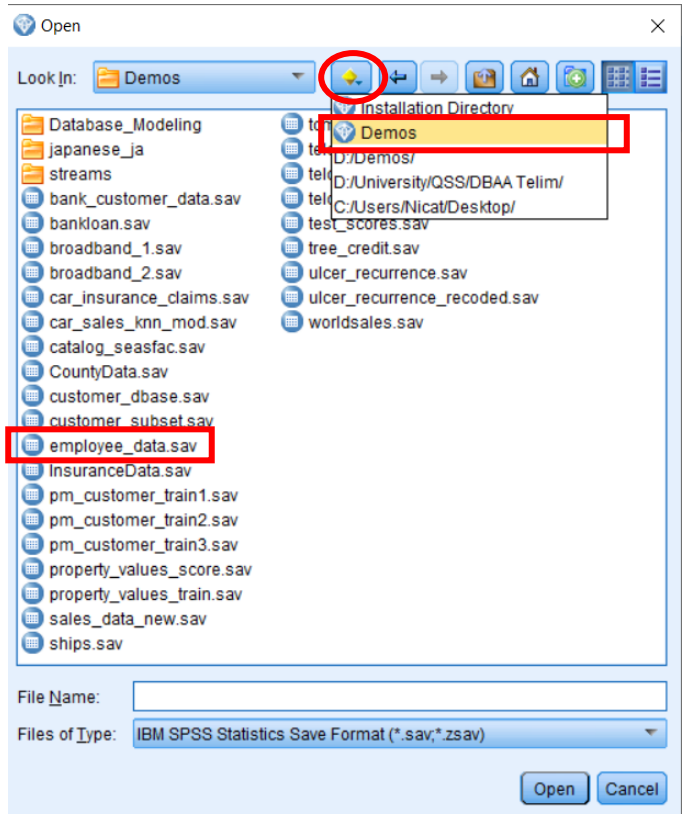


Bütün hüquqlar qorunur.

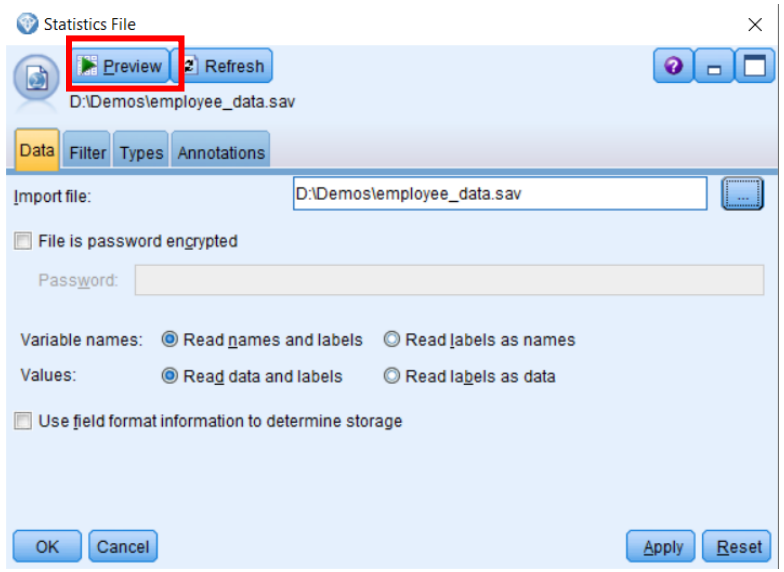
Connecting Data

4. Click **Yellow Diamond (Nugget)** button and select **Demos**;
5. Double click on the *"employee_data.sav"*;
6. Click **Preview**;
7. Click **Ok** .

3



4



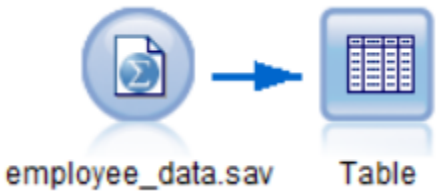
Bütün hüquqlar qorunur.

Table node

1. Double click on the **Table** from **Output** tab;
2. Click **Run** ( ) button from the toolbar.

1

2



Bütün hüquqlar qorunur.

Table node output

Table (10 fields, 474 records)

File Edit Generate

Table Annotations

	ID	GENDER	BDATE	EDUC	JOBCAT	SALARY	SALBEGIN	JOBTIME	PREVEXP	MINORITY
1	1.0...	m	1952-02-03	15.000	3.000	57000....	27000.000	98.000	144.000	0.000
2	2.0...	m	1958-05-23	16.000	1.000	40200....	18750.000	98.000	36.000	0.000
3	3.0...	f	1929-07-26	12.000	1.000	21450....	12000.000	98.000	381.000	0.000
4	4.0...	f	1947-04-15	8.000	1.000	21900....	13200.000	98.000	190.000	0.000
5	5.0...	m	1955-02-09	15.000	1.000	45000....	21000.000	98.000	138.000	0.000
6	6.0...	m	1958-08-22	15.000	1.000	32100....	13500.000	98.000	67.000	0.000
7	7.0...	m	1956-04-26	15.000	1.000	36000....	18750.000	98.000	114.000	0.000
8	8.0...	f	1966-05-06	12.000	1.000	21900....	9750.000	98.000	0.000	0.000
9	9.0...	f	1946-01-23	15.000	1.000	27900....	12750.000	98.000	115.000	0.000
10	10....	f	1946-02-13	12.000	1.000	24000....	13500.000	98.000	244.000	0.000
11	11....	f	1950-02-07	16.000	1.000	30300....	16500.000	98.000	143.000	0.000
12	12....	m	1966-01-11	8.000	1.000	28350....	12000.000	98.000	26.000	1.000
13	13....	m	1960-07-17	15.000	1.000	27750....	14250.000	98.000	34.000	1.000
14	14....	f	1949-02-26	15.000	1.000	35100....	16800.000	98.000	137.000	1.000
15	15....	m	1962-08-29	12.000	1.000	27300....	13500.000	97.000	66.000	0.000
16	16....	m	1964-11-17	12.000	1.000	40800....	15000.000	97.000	24.000	0.000
17	17....	m	1962-07-18	15.000	1.000	46000....	14250.000	97.000	48.000	0.000
18	18....	m	1956-03-20	16.000	3.000	10375...	27510.000	97.000	70.000	0.000
19	19....	m	1962-08-19	12.000	1.000	42300....	14250.000	97.000	103.000	0.000
20	20....	f	1940-01-23	12.000	1.000	26250....	11550.000	97.000	48.000	0.000
21	21....	f	1963-02-19	16.000	1.000	38850....	15000.000	97.000	17.000	0.000

OK

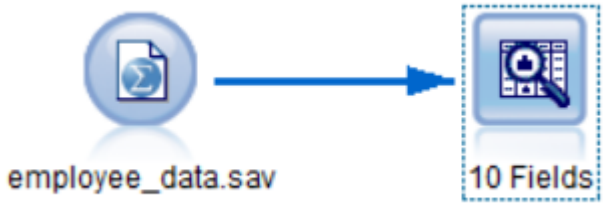
Bütün hüquqlar qorunur.

Data Audit Node

1. Open *“employee_data.sav”*
2. Double click the **Data Audit** node from **Field Ops** tab;
3. Click **Run** ( ) button from the toolbar;

1

2



Bütün hüquqlar qorunur.

Data Audit Node Output

Data Audit of [10 fields]

File Edit Generate

Audit Quality Annotations

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
ID		Continuous	1.000	474.000	237.500	136.976	0.000	--	474
GENDER		Nominal	--	--	--	--	--	2	474
BDATE		Continuous	1929-02-10	1971-02-10	--	--	--	--	473
EDUC		Ordinal	8.000	21.000	--	--	--	10	474
JOBCAT		Ordinal	1.000	3.000	--	--	--	3	474
SALARY		Continuous	15750.000	135000.000	34419.568	17075.661	2.125	--	474
SALBE...		Continuous	9000.000	79980.000	17016.086	7870.638	2.853	--	474
JOBTIME		Continuous	63.000	98.000	81.110	10.061	-0.053	--	474
PREVE...		Continuous	0.000	476.000	95.861	104.586	1.510	--	474
MINORI...		Ordinal	0.000	1.000	--	--	--	2	474

[1] Indicates a multimode result [2] Indicates a sampled result

OK

Bütün hüquqlar qorunur.

Thank you!

Data Science Academy