AGENDA

1. Comparing Means: ANOVA Test
2. Data Understanding
3. Data Exploration, Visualization

Data Science Academy

I

Comparing Means: ANOVA Test

Data Science Academy

One Sample T Test

- The one sample t test compares the mean score of our sample with a known value, usually with the population mean. The sample mean is the observed average, while the population mean is the expected average.
- This test is useful when we want to know whether our sample comes from a particular population.
- If Sig.<0,05 means sample statistics is significant for population mean.

Assumptions:

1. The sample units are randomly extracted from the population.
2. Our variable is normally distributed.
3. There are no significant outliers in our data.

What to report:

- ○ The p value (Sig. column)
- ○ The mean difference between the sample mean and the population mean

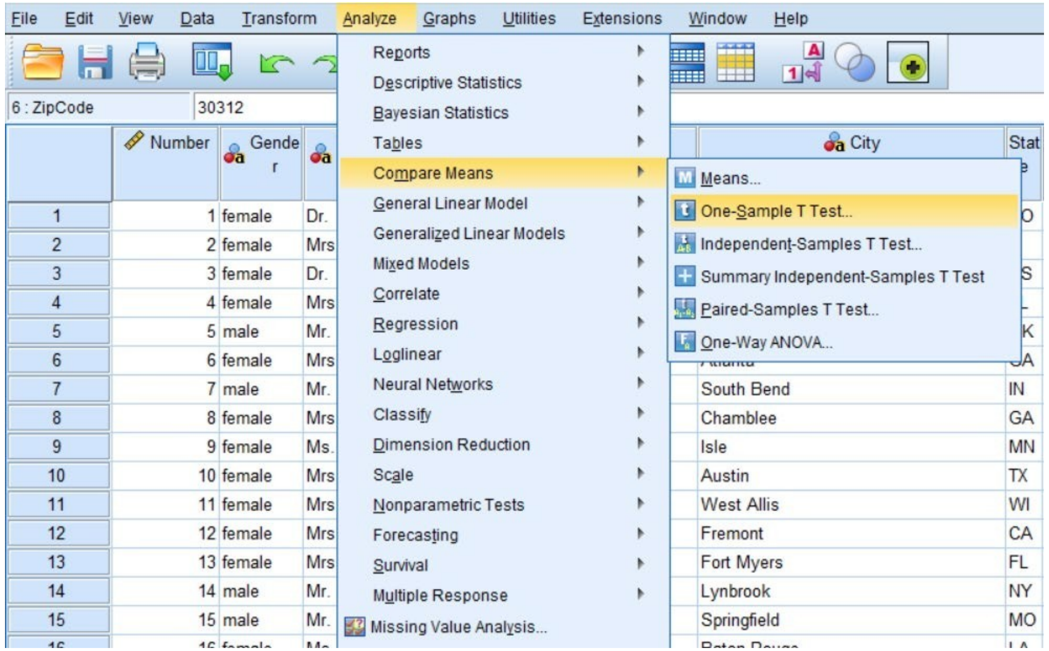
Using SPSS for One Sample T-test

According to the US Department of Health and Human Services, average height of Americans is 175.3 cm. Let’s test this for our datasets.

Using SPSS for One Sample T-test

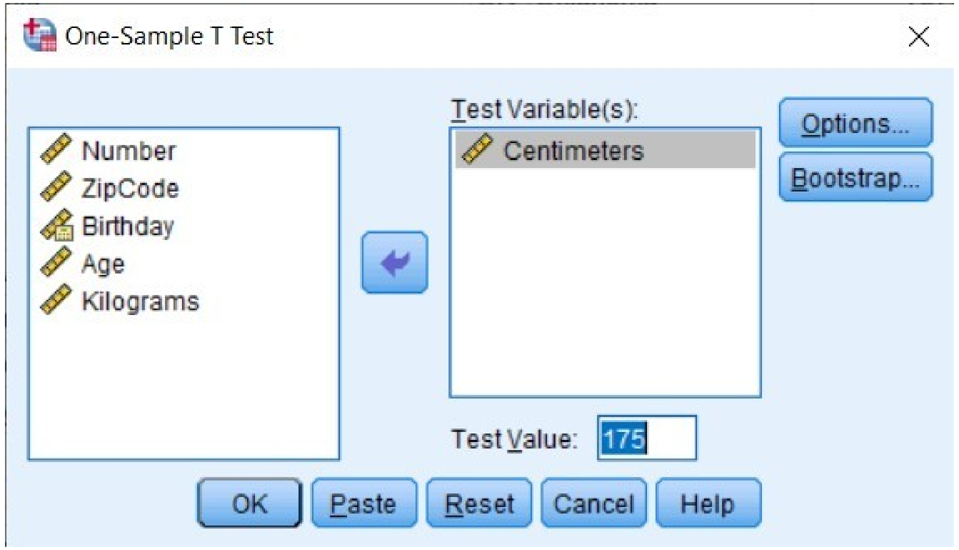
1. Open “US-Insurance-Clients.csv”
2. Select Compare Means form Analyze tab
3. Select One-Sample T Test option

1. Enter “Centimeters” to the Test Variable(s)
2. Write “ 175” to the Test Value
3. Click OK.



6 : ZipCode 30312

	Number	Gender		City	State
1	1	female	Dr.		O
2	2	female	Mrs		
3	3	female	Dr.		S
4	4	female	Mrs		
5	5	male	Mr.		K
6	6	female	Mrs	Atlanta	GA
7	7	male	Mr.	South Bend	IN
8	8	female	Mrs	Chamblee	GA
9	9	female	Ms.	Isle	MN
10	10	female	Mrs	Austin	TX
11	11	female	Mrs	West Allis	WI
12	12	female	Mrs	Fremont	CA
13	13	female	Mrs	Fort Myers	FL
14	14	male	Mr.	Lynbrook	NY
15	15	male	Mr.	Springfield	MO
16	16	female	Ms	Baton Rouge	LA

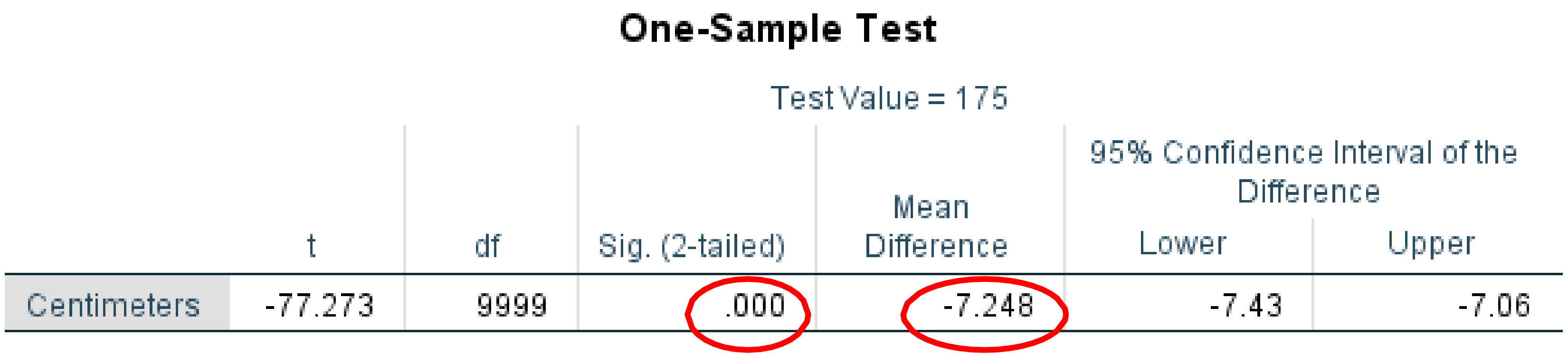


Test Variable(s): Centimeters

Test Value: 175

Using SPSS for One Sample T-test

- Since $p < 0.05$, we conclude that the mean height of the sample is significantly different than the average height of the overall adult population.
- The mean difference between the "observed" sample mean and the population mean is -7.248. The negative mean difference in this example indicates that the mean height of the sample is less than the population mean (175).



One-Sample Test

	Test Value = 175					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Centimeters	-77.273	9999	.000	-7.248	-7.43	-7.06

Independent and Paired Sample T-test

○ The paired sample t test determines mean difference between two variables, measured on the same subjects, at two different moments, is statistically significant.

○ In order to run the paired-Samples t test, we must have two paired measurements for each subject.

Tests	Bravo	Araz
Bread	0.3	0.4
Apple	1.0	1.2
Lays	2.5	2.3
Water	0.25	0.3
…	…	

○ The independent sample t test finds out difference between the means of two independent groups on a continuous variable. More precisely, this test lets us determine if the difference between the means is statistically significant.

○ If the p value is lower than 0,05, there is significant difference between population means.

Gender	Monthly Income
Female	131160
Female	41890
Male	193280
Male	83210
…	…

Assumptions:

Paired:

1. The differences between the scores of the variables are normally distributed.
2. The differences between the scores of the variables do not present significant outliers.

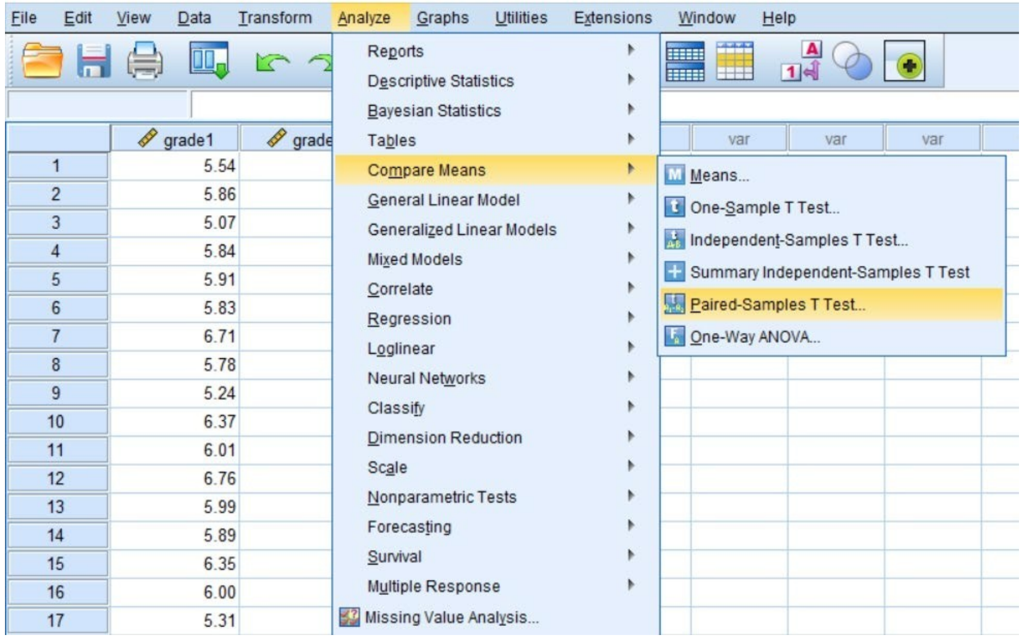
Independent

1. There is independence of observations; in other words, there is no connection between the observations(measurements) in the two groups.
2. The dependent variable is normally distributed in both groups.
3. The dependent variable has no significant outliers in either group.
4. The variances of the dependent variable in the two groups are equal (in other words, we have homogeneity of variances).

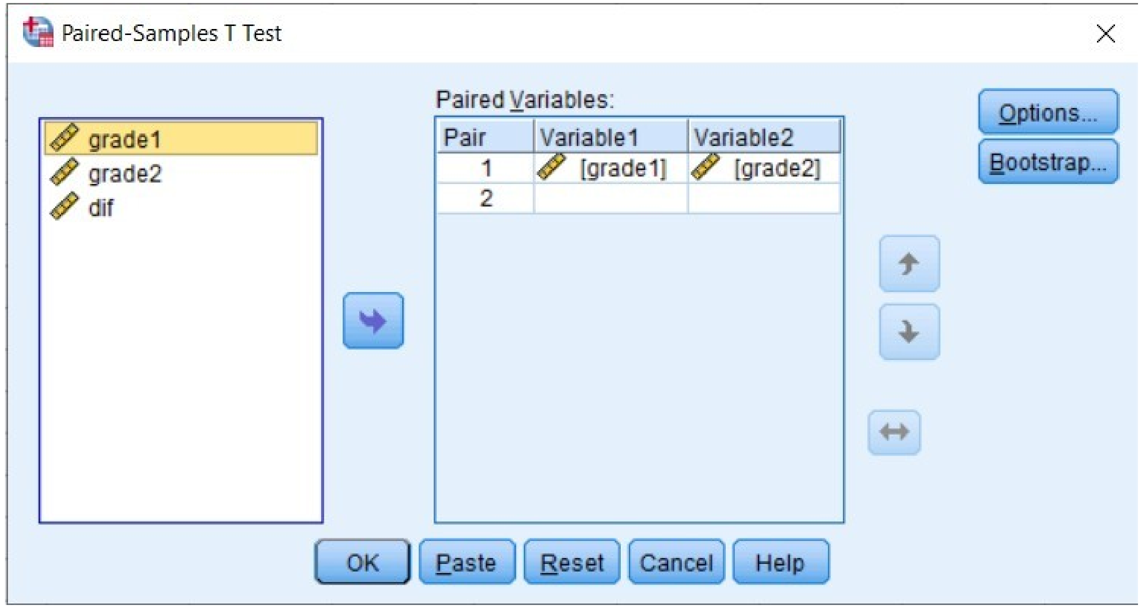
Using SPSS for Paired-Samples T test

1. Open “math-test.sav”
2. Select Compare Means form Analyze tab
3. Select Paired Sample T-test option

1. Enter Grade1 and Grade2 to the Variables
2. Click OK.



	grade1	grade
1	5.54	
2	5.86	
3	5.07	
4	5.84	
5	5.91	
6	5.83	
7	6.71	
8	5.78	
9	5.24	
10	6.37	
11	6.01	
12	6.76	
13	5.99	
14	5.89	
15	6.35	
16	6.00	
17	5.31	

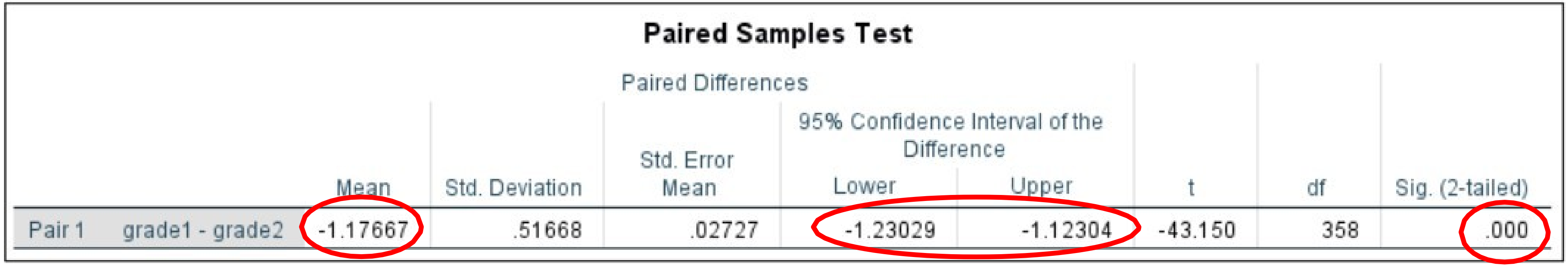


Paired Variables:

Pair	Variable1	Variable2
1	[grade1]	[grade2]
2		

Using SPSS for Paired-Samples T test

- There is a significant difference between population means of grades (p < 0.05).
- On average, grade1 is -1.176 points less than grade2 (95% Confidence Interval of difference [-1.23, -1.12]).

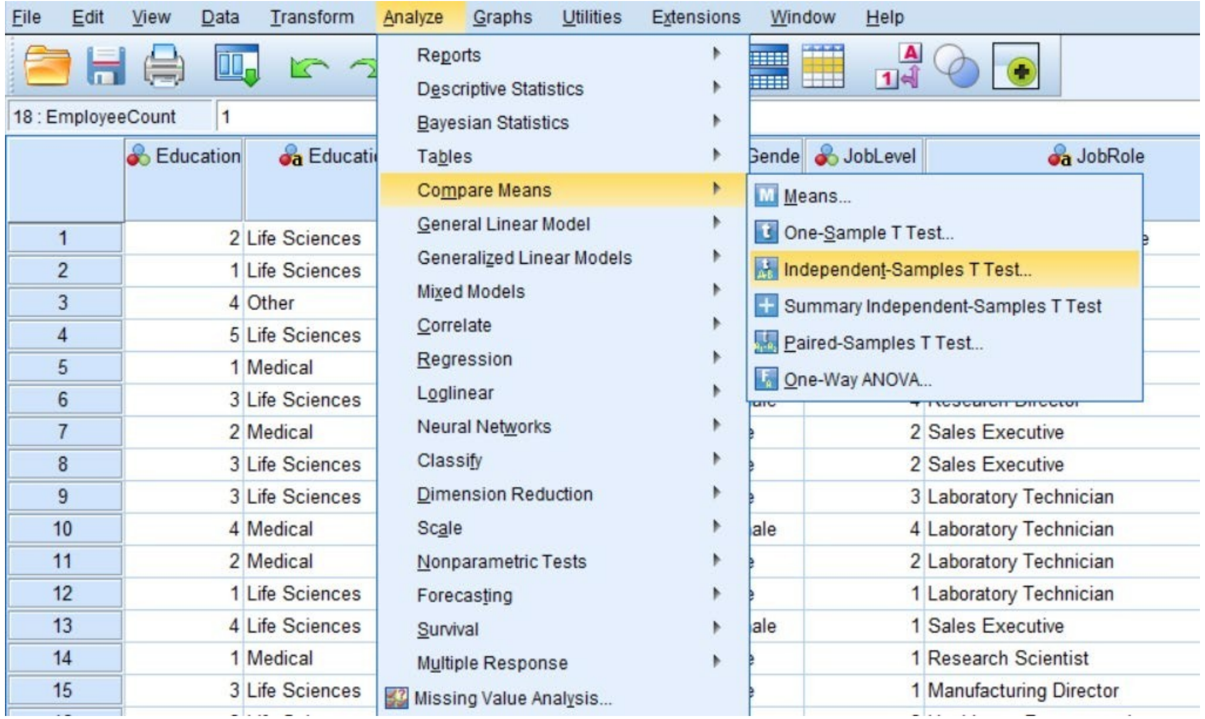


Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	grade1 - grade2	-1.17667	.51668	.02727	-1.23029	-1.12304	-43.150	358	.000

Using SPSS for Independent Samples T Test

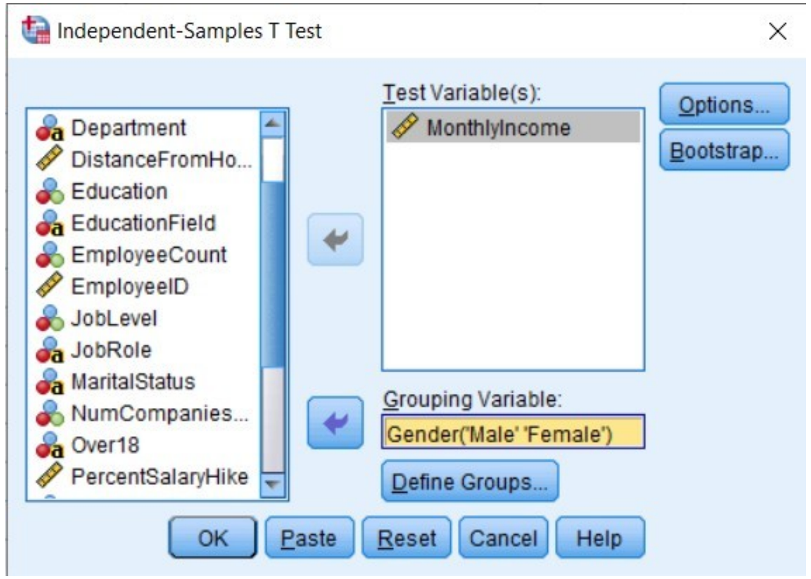
- Open “general_data.csv”
- Select Compare Means form Analyze tab;
- Select Independent-Sample T-test option;



	Education	Educati	Gende	JobLevel	JobRole
1	2	Life Sciences			
2	1	Life Sciences			
3	4	Other			
4	5	Life Sciences			
5	1	Medical			
6	3	Life Sciences			
7	2	Medical		2	Sales Executive
8	3	Life Sciences		2	Sales Executive
9	3	Life Sciences		3	Laboratory Technician
10	4	Medical	ale	4	Laboratory Technician
11	2	Medical		2	Laboratory Technician
12	1	Life Sciences		1	Laboratory Technician
13	4	Life Sciences	ale	1	Sales Executive
14	1	Medical		1	Research Scientist
15	3	Life Sciences		1	Manufacturing Director

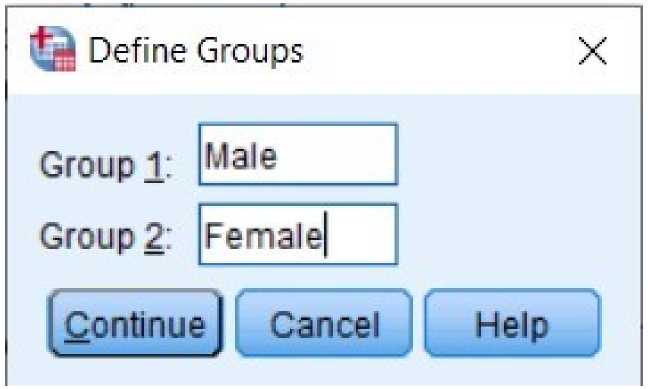
Using SPSS for Independent-Samples T Test

- Enter “MonthlyIncome” to Test Variables(s) and “Gender” to the Grouping Variable(s);
- Click Define Groups;
- Write “Male” to Group 1 and “Female” to Group 2;
- Click Continue and click OK.


Independent-Samples T Test

Test Variable(s): MonthlyIncome

Grouping Variable: Gender('Male' 'Female')

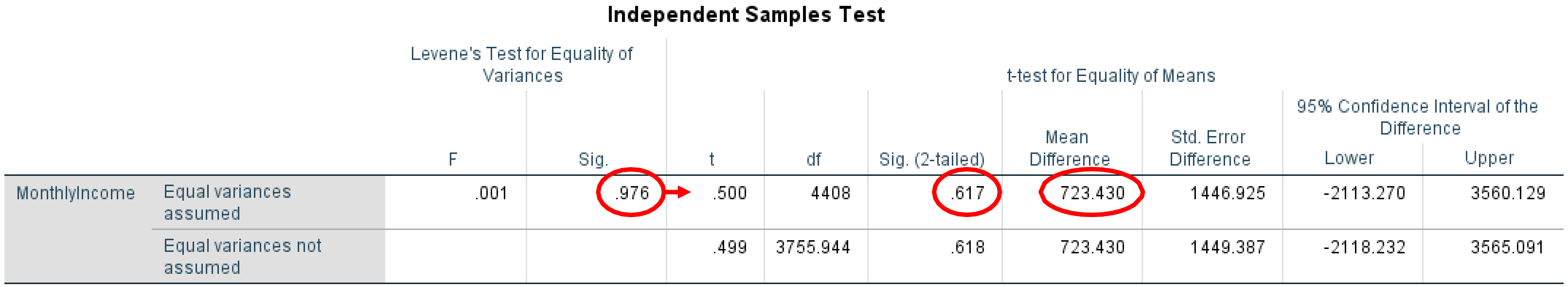

Define Groups

Group 1: Male

Group 2: Female

Using SPSS for Independent-Samples T Test

- The p-value of Levene's test is printed as ".976", so we conclude that the variance in monthly income of males' is not significantly different than that of females. This tells us that we should look at the "Equal variances assumed" row for the t test (and corresponding confidence interval) results. (If this test result had been significant -- that is, if we had observed p < α -- then we would have used the "Equal variances not assumed" output.)
- There is not significant difference between average monthly income of males and females (p > 0.05).
- On average, monthly income of male is 723.4 dollar higher than that of female but this difference is not significant for population.



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
MonthlyIncome	Equal variances assumed	.001	.976	.500	4408	.617	723.430	1446.925	-2113.270	3560.129
	Equal variances not assumed			.499	3755.944	.618	723.430	1449.387	-2118.232	3565.091

One-Way ANOVA

The one-way analysis of variance (or ANOVA) can help us determine whether there are significant differences between the means of three or more groups, for a continuous variable.

In order to use this test, we must have two variables:

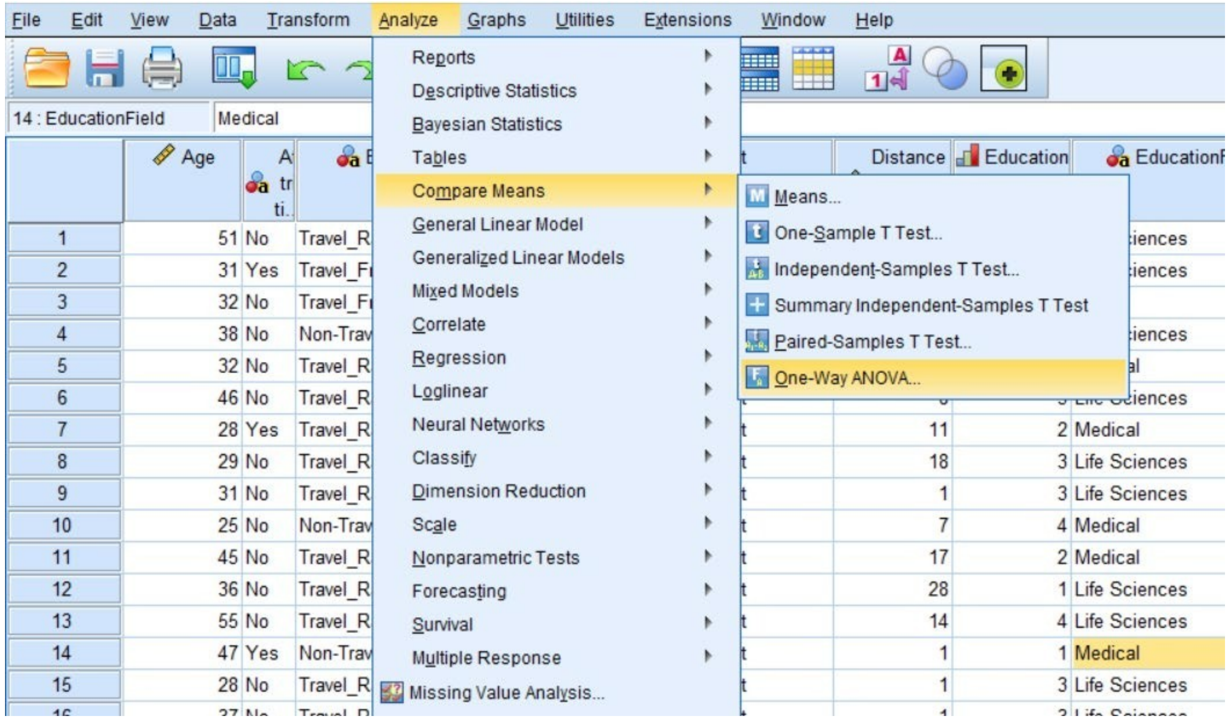
- a categorical independent variable with three or more categories
- a continuous dependent variable.

One-Way ANOVA

Assumptions:

- The independent variable is categorical, with three or more categories.
- The dependent variable is continuous. The dependent variable is normally distributed in all groups.
- The dependent variable does not present significant outliers in any group.
- The dependent variable has equal variances in all groups. If this condition is not met, we have to use a robust version of the F test, called the Welch test.

Using SPSS for One-Way ANOVA

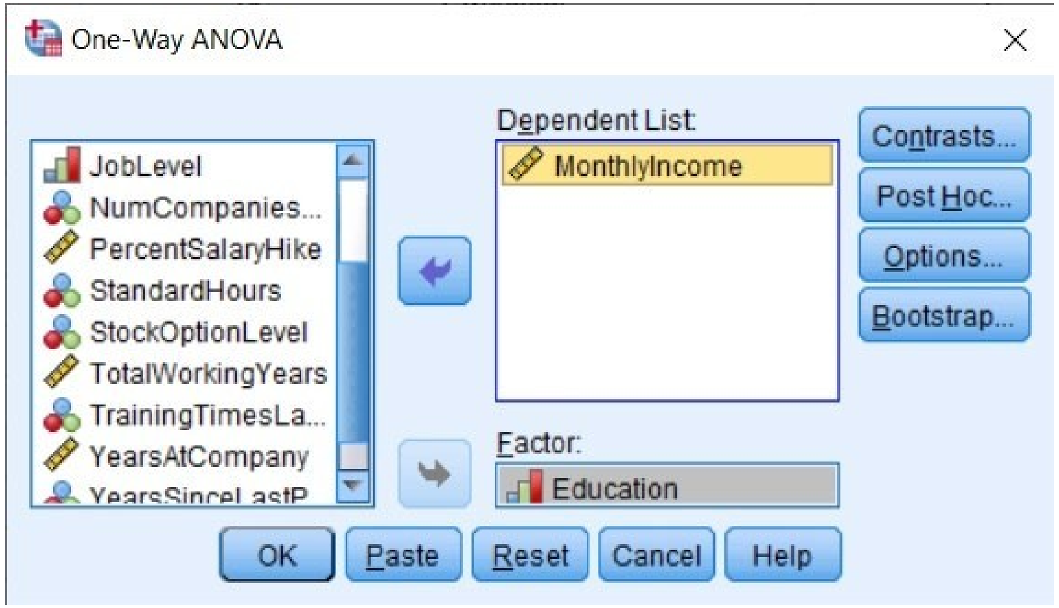


- Select Compare Means form Analyze tab;
- Select One-Way ANOVA option;

Using SPSS for One-Way ANOVA

1. Add “MonthlyIncome” to Dependent List and “Education” to Factor.
2. Click Options.

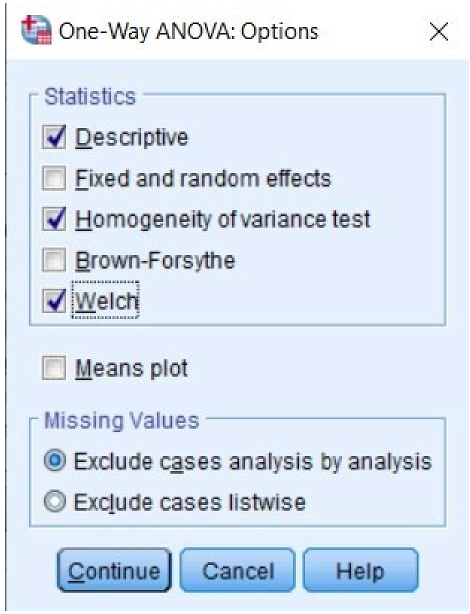
1. Select “Descriptive”, “Homogeneity of variance test” and “Welch”.
2. Click Continue.



One-Way ANOVA

Dependent List:
MonthlyIncome

Factor:
Education



One-Way ANOVA: Options

Statistics
☑ Descriptive
☐ Fixed and random effects
☑ Homogeneity of variance test
☐ Brown-Forsythe
☑ Welch

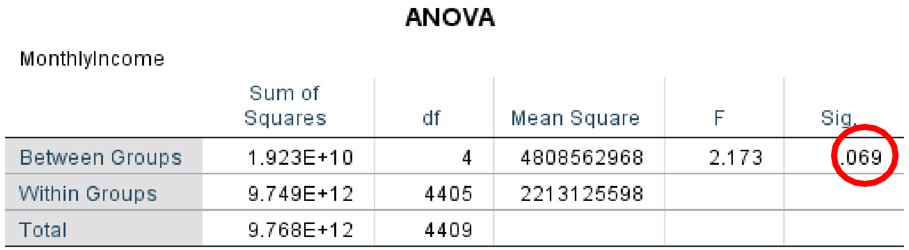
☐ Means plot

Missing Values
● Exclude cases analysis by analysis
○ Exclude cases listwise

Using SPSS for One-Way ANOVA

- We have not a significant result. p-value of .069 (which is greater than the .05 alpha level). This means there is not a statistically significant difference between the means of the different levels of the education.

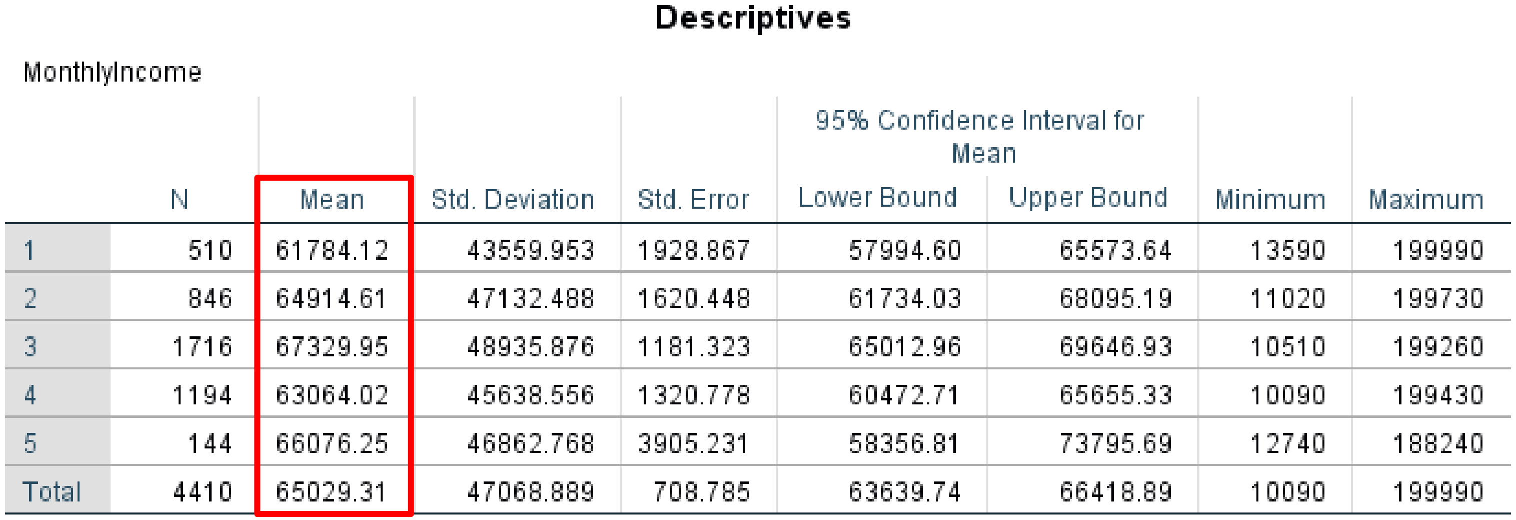
- You can see that our sample data produces not significant difference in the mean scores of the five levels of our education variable.



ANOVA

MonthlyIncome

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.923E+10	4	4808562968	2.173	.069
Within Groups	9.749E+12	4405	2213125598		
Total	9.768E+12	4409			



Descriptives

MonthlyIncome

					95% Confidence Interval for Mean			
	N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
1	510	61784.12	43559.953	1928.867	57994.60	65573.64	13590	199990
2	846	64914.61	47132.488	1620.448	61734.03	68095.19	11020	199730
3	1716	67329.95	48935.876	1181.323	65012.96	69646.93	10510	199260
4	1194	63064.02	45638.556	1320.778	60472.71	65655.33	10090	199430
5	144	66076.25	46862.768	3905.231	58356.81	73795.69	12740	188240
Total	4410	65029.31	47068.889	708.785	63639.74	66418.89	10090	199990

II

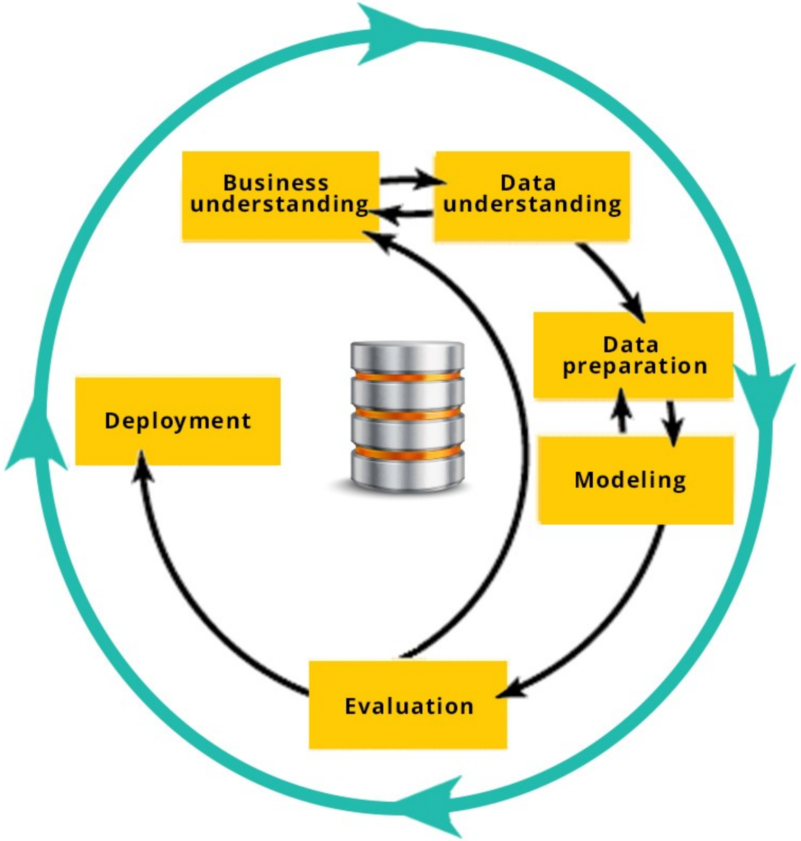
Data Understanding

Data Science Academy

CRISP-DM

- CRISP-DM is an industry-proven process model provides an overview of the data mining life cycle.

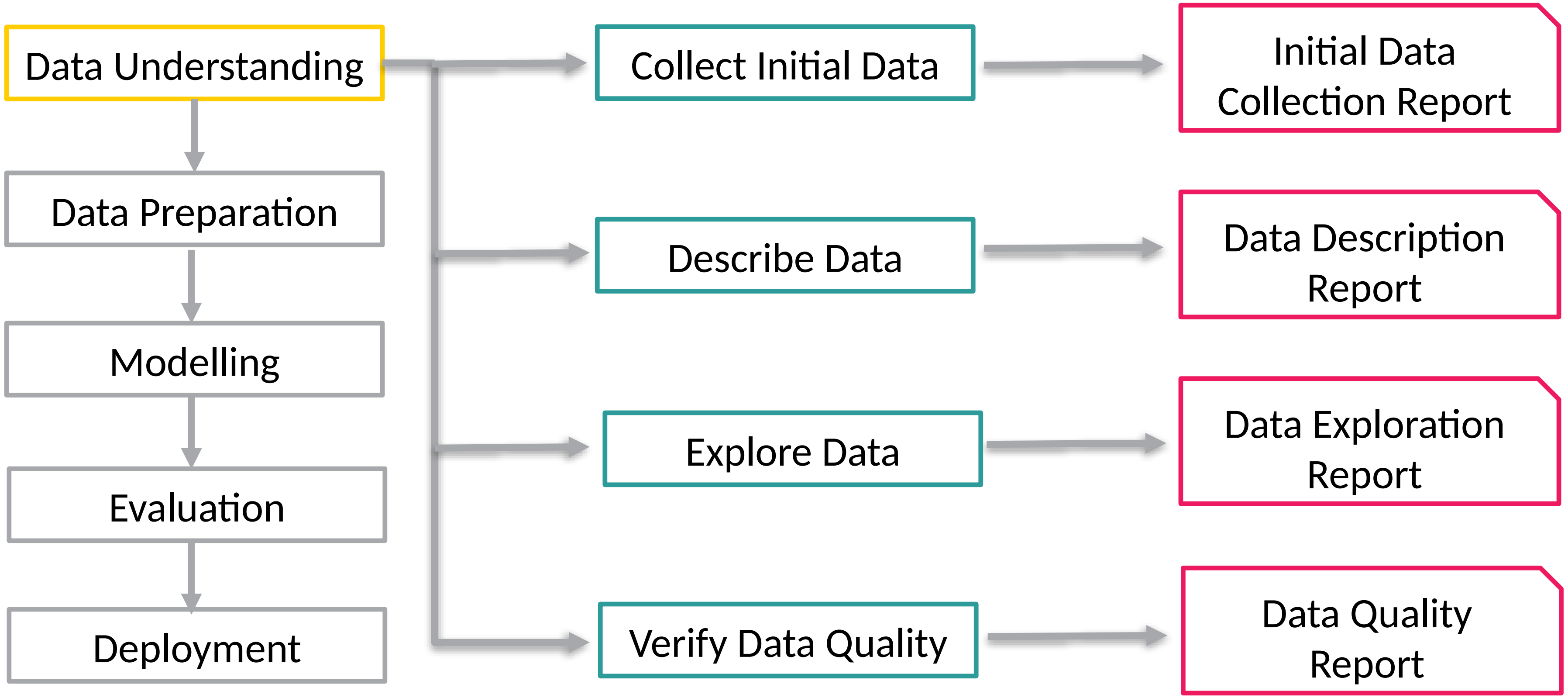
- The life cycle model consists of six phases. As a **methodology**, it includes descriptions of phases of the project, the tasks involved with each phase, and an explanation of the relationships between these tasks.



Cross-Industry Standard Process for Data Mining

Data Understanding

- Data understanding involves accessing the data and exploring it using tables and graphics that can be organized in SPSS Modeler using the CRISP-DM project tool. This enables you to determine the quality of the data and describe the results of these steps in the project documentation.



Collecting Initial Data

- At this point in CRISP-DM, you're ready to access data and bring it into IBM SPSS Modeler. Data come from a variety of sources, such as:

  - **Existing data.** This includes a wide variety of data, such as transactional data, survey data, Web logs, etc. Consider whether the existing data are enough to meet your needs.

  - **Purchased data.** Does your organization use supplemental data, such as demographics? If not, consider whether it may be needed.

  - **Additional data.** If the above sources don't meet your needs, you may need to conduct surveys or begin additional tracking to supplement the existing data stores.

Collecting Initial Data

- Which attributes (columns) from the database seem most promising?
- Which attributes seem irrelevant and can be excluded?
- Is there enough data to draw generalizable conclusions or make accurate predictions?
- Are there too many attributes for your modeling method of choice?
- Are you merging various data sources? If so, are there areas that might pose a problem when merging?
- Have you considered how missing values are handled in each of your data sources?

Collecting Initial Data
E-Retail Example

- The e-retailer in this example uses several important data sources, including:
  - **Web logs.** The raw access logs contain all of the information on how customers navigate the Web site.
  - **Purchase data.** When a customer submits an order, all of the information pertinent to that order is saved.
  - **Product database.** The product attributes may be useful when determining "related" products.
  - **Customer database.** This database contains extra information collected from registered customers.

Data Collection Report

- Using the material gathered in the previous step, write a data collection report.
- The report can be added to the project Web site or distributed to the team.
- It can also be combined with the reports prepared in the next steps--data description, exploration, and quality verification.
- These reports will guide your work throughout the data preparation phase.

Collecting Initial Data

- At this point in CRISP-DM, you’re ready to access data and bring it into SPSS Modeler.

Variable File - node imports data from free-field text files.

Excel node - imports data from MS Excel in the .xlsx file format.

Statistics File - node imports data from SPSS Statistics in the .sav, .zsav file format.

Data Description

There are many ways to describe data, but most descriptions focus on the quantity and quality of the data — how much data is available and the condition of the data.

- ○ Listed below are some key characteristics to address when describing data.

    - Amount of data
    - Value types
    - Coding Schemes
    - Statistical indicators
    - Distributions

Writing a Data Description Report

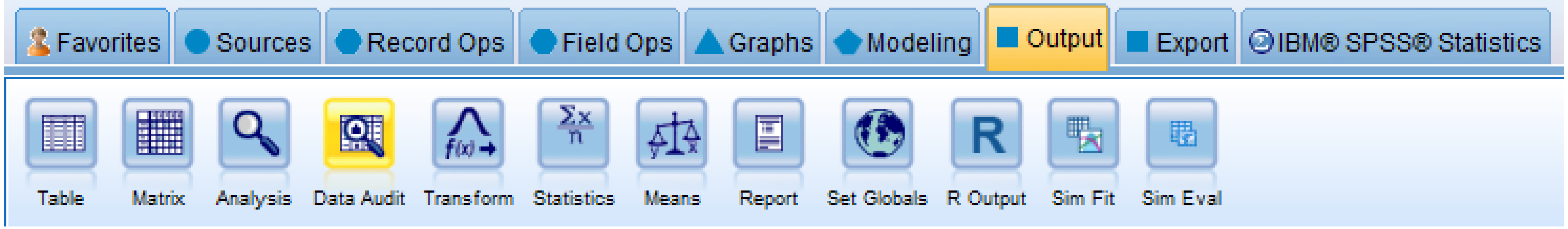
- **Data Quantity**
  - What is the format of the data?
  - Identify the method used to capture the data--for example, ODBC.
  - How large is the database (in numbers of rows and columns)?
- **Data Quality**
  - Does the data include characteristics relevant to the business question?
  - What data types are present (symbolic, numeric, etc.)?
  - Did you compute basic statistics for the key attributes? What insight did this provide into the business question?
  - Are you able to prioritize relevant attributes? If not, are business analysts available to provide further insight?

Writing a Data Description Report

- What sort of hypotheses have you formed about the data?
- Which attributes seem promising for further analysis?
- Have your explorations revealed new characteristics about the data?
- How have these explorations changed your initial hypothesis?
- Can you identify particular subsets of data for later use?
- Take another look at your data mining goals. Has this exploration altered the goals?

Statistics indicators

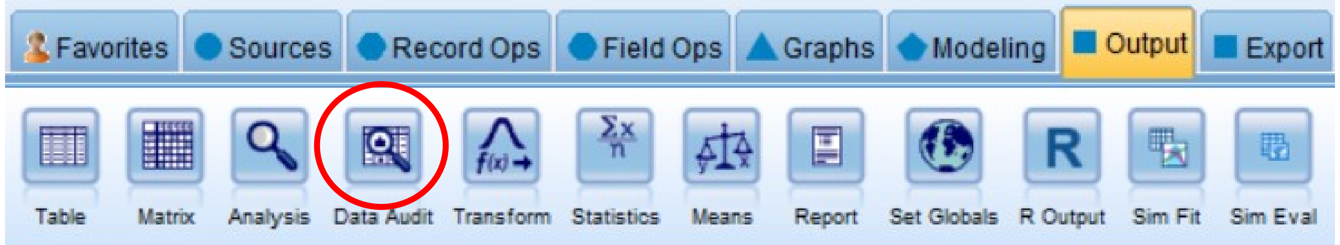
- Statistics indicators consist of minimum, maximum, mean, median, mode, standart deviation, variance, skewness and etc.
- There are several ways we can get this information in SPSS Modeler. But most common method is to use **Data Audit node** in **Output** palette.
- The Data Audit node provides a comprehensive first look at the data you bring into SPSS Modeler, presented in an easy-to-read matrix that can be sorted and used to generate full-size graphs and a variety of data preparation nodes.



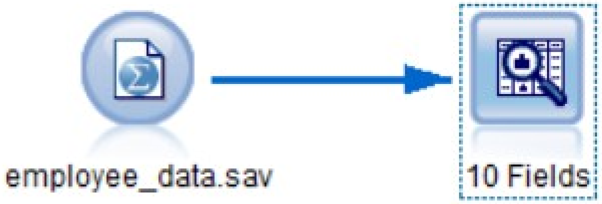
Data Audit Node

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

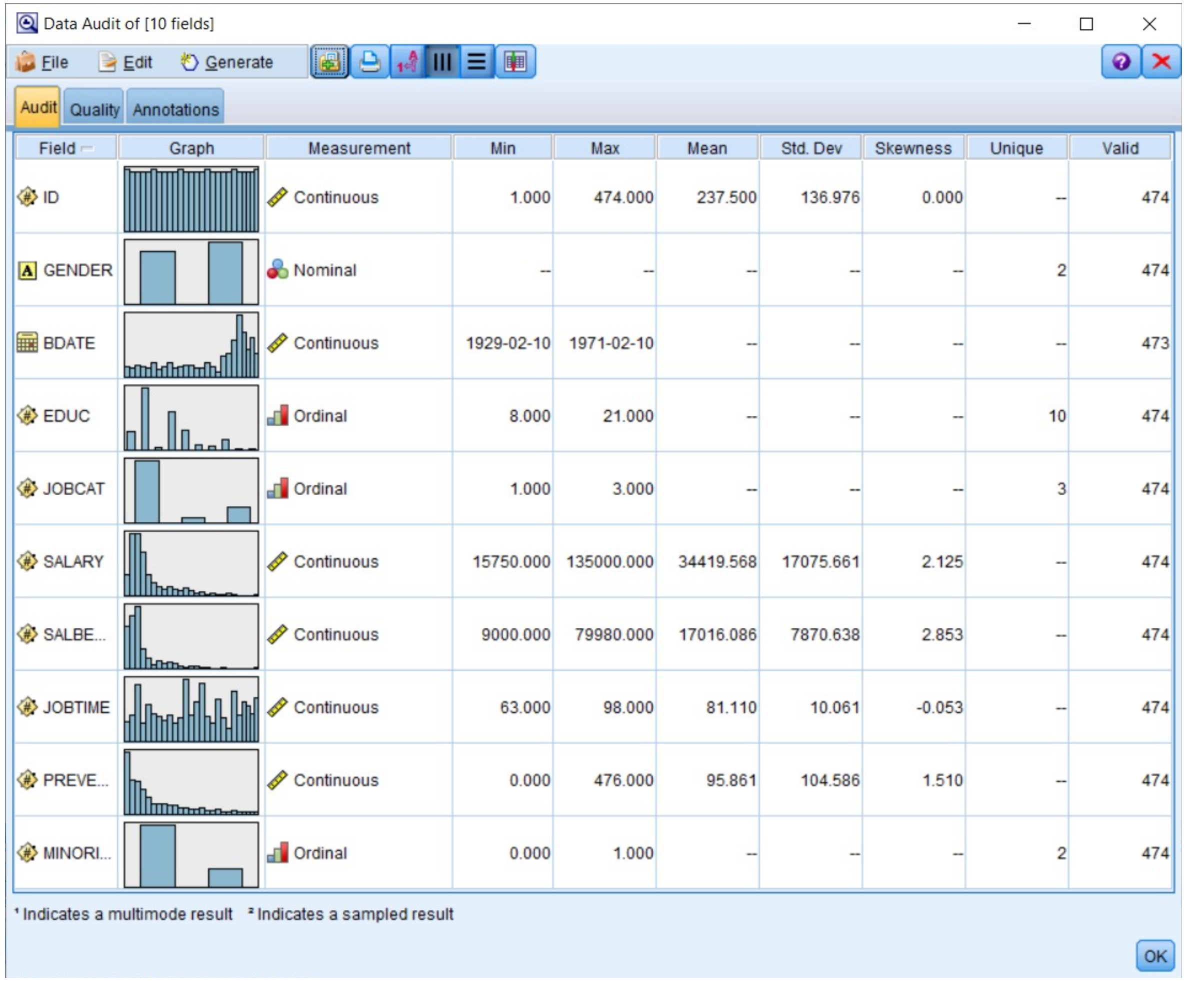


www.dsa.az

Bütün hüquqlar qorunur.

Data Audit Node Output

- After running node, for each field you can get:
  - some basic statistical indicators;
  - graphs
  - measurements



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

III

Data Exploration, Visualization

Data Science Academy

Data Exploration

- Explore the data with the **tables, charts, and other visualization tools**

- Formulate hypotheses and shape the data transformation tasks that take place during data preparation.

Data Exploration Report

- The report includes first findings or initial hypothesis and their impact on the remainder of the project.

- The report may also include graphs and plots that indicate data characteristics or point to interesting data subsets worthy of further examination.

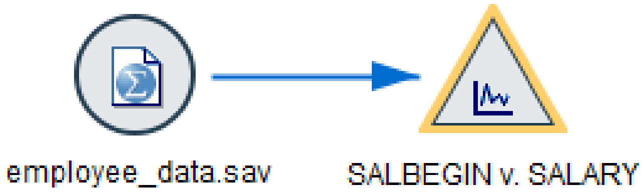
Plot node

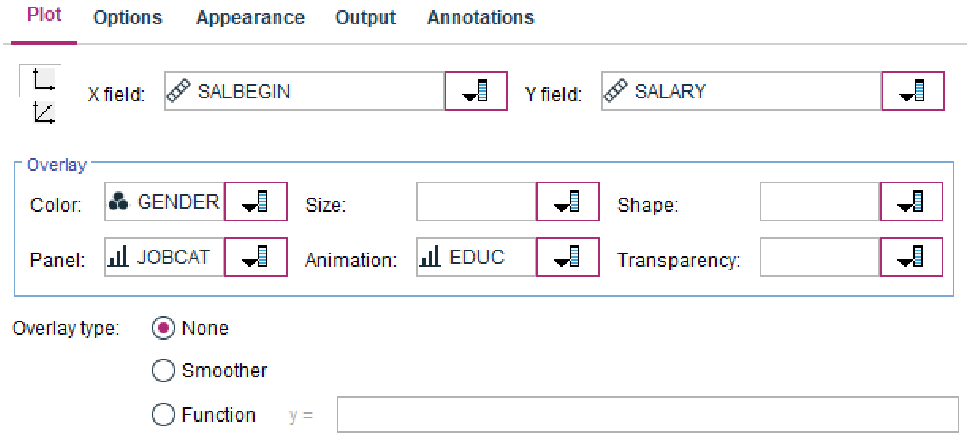
Plot node shows the relationship between numeric fields. You can create a plot using points (also known as a scatterplot), or you can use lines.

1. Add **Statistics file** from **Source** palette;
2. Import *"employee_data.sav";*
3. Add **Plot** node from **Graphs** palette;
4. Add *"SALBEGIN"* to **X field** and

1. *"SALARY"* to **Y field**;

2. Add *"GENDER"* to **Color**, *"EDUC"* to **Animation** and *"JOBCAT"* to **Panel**;
3. Click Run.





X field: SALBEGIN

Y field: SALARY

Color: GENDER

Size:

Shape:

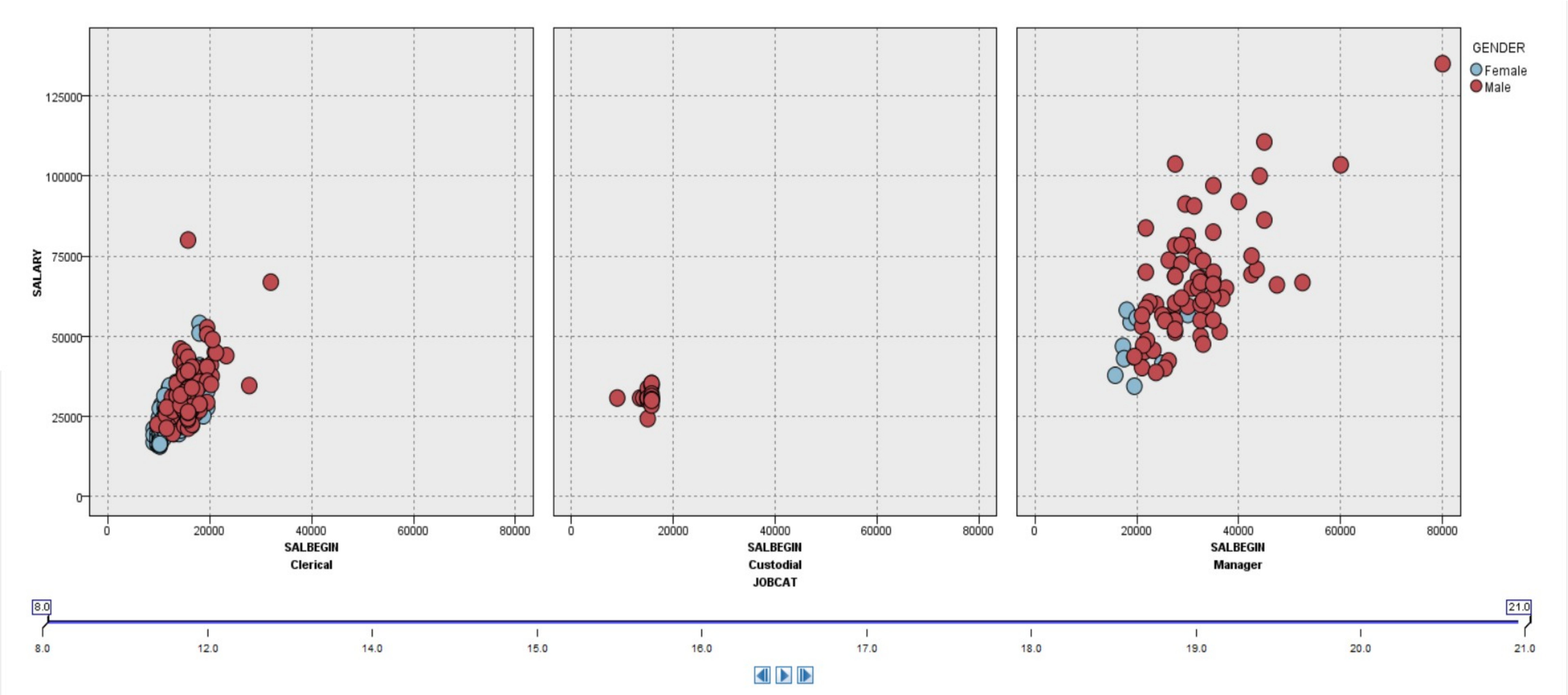
Panel: JOBCAT

Animation: EDUC

Transparency:

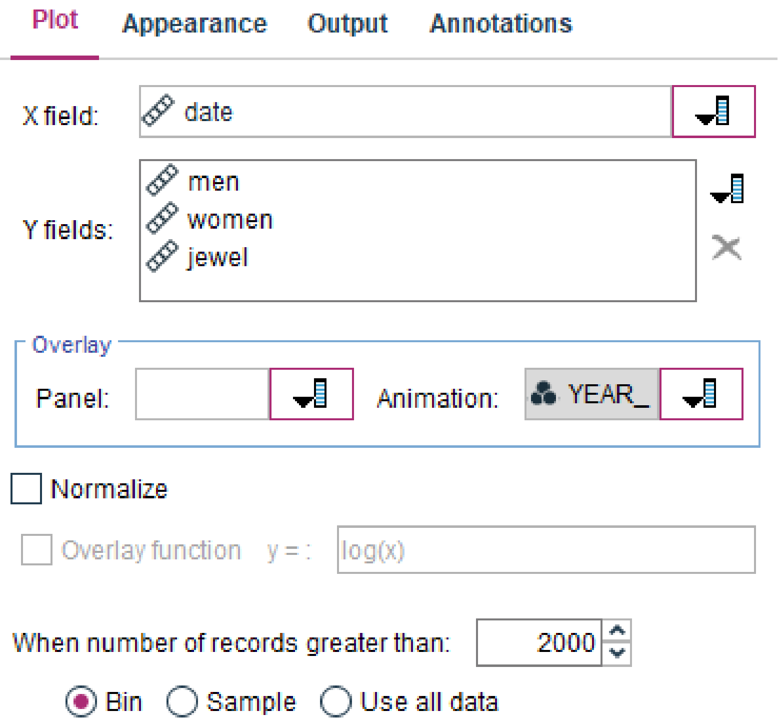
Overlay type: ● None ○ Smoother ○ Function y =

Output


SALARY
125000
100000
75000
50000
25000
0
0
20000
40000
60000
80000
SALBEGIN
Clerical
SALBEGIN
Custodial
JOBCAT
SALBEGIN
Manager
GENDER
Female
Male
8.0
21.0
8.0
12.0
14.0
15.0
16.0
17.0
18.0
19.0
20.0
21.0

Multiplot node

1. Add **Statistics file** from **Source** palette;
2. Import *"catalog_seafac.sav"*;
3. Add **Multiplot** node from **Graphs** palette;
4. Add *"date"* to **X field** and *"men"*, *"women"*, *"jewel"* to **Y field**;
5. Add *"YEAR_"* to **Animation**;
6. Click Run.



X field: date

Y fields: men, women, jewel

Overlay

Panel:

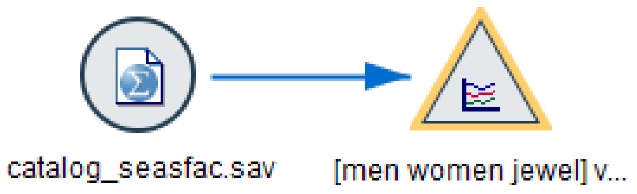
Animation: YEAR_

□ Normalize

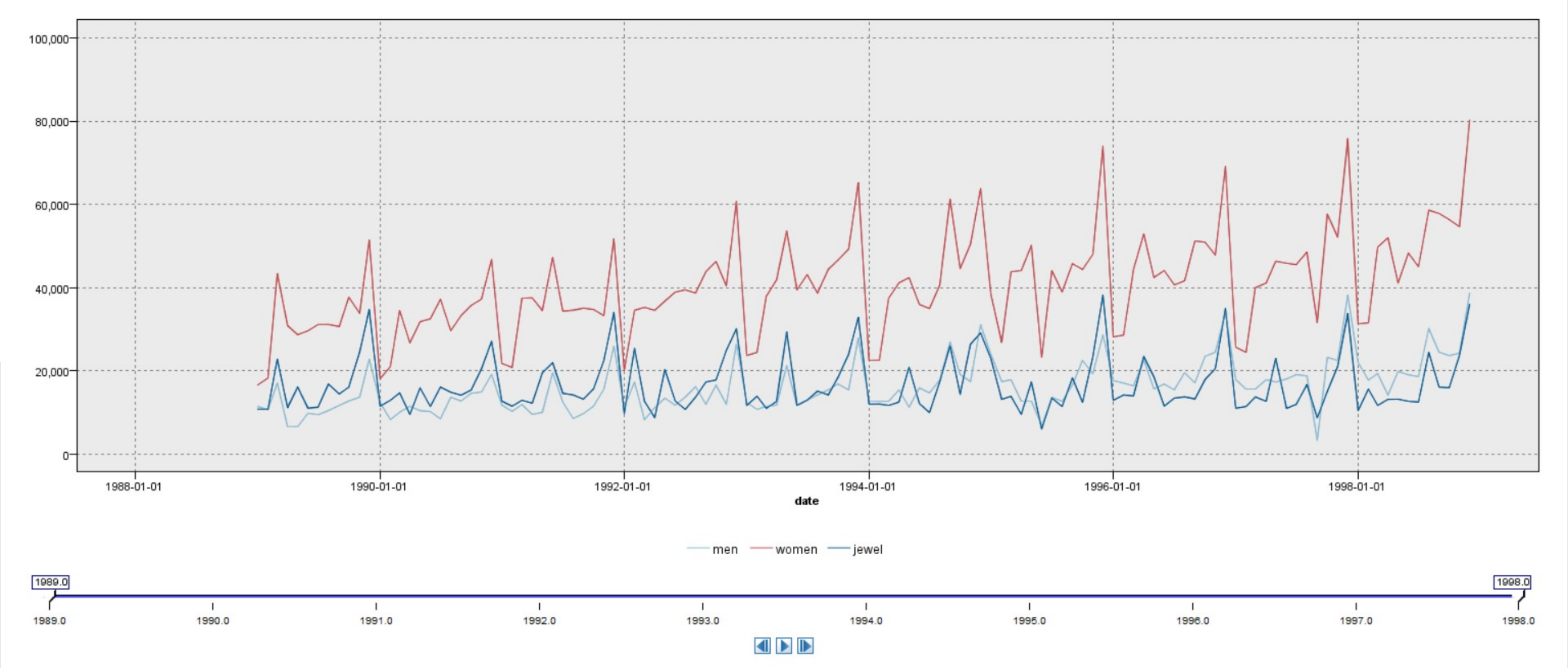
□ Overlay function y = : log(x)

When number of records greater than: 2000

● Bin ○ Sample ○ Use all data

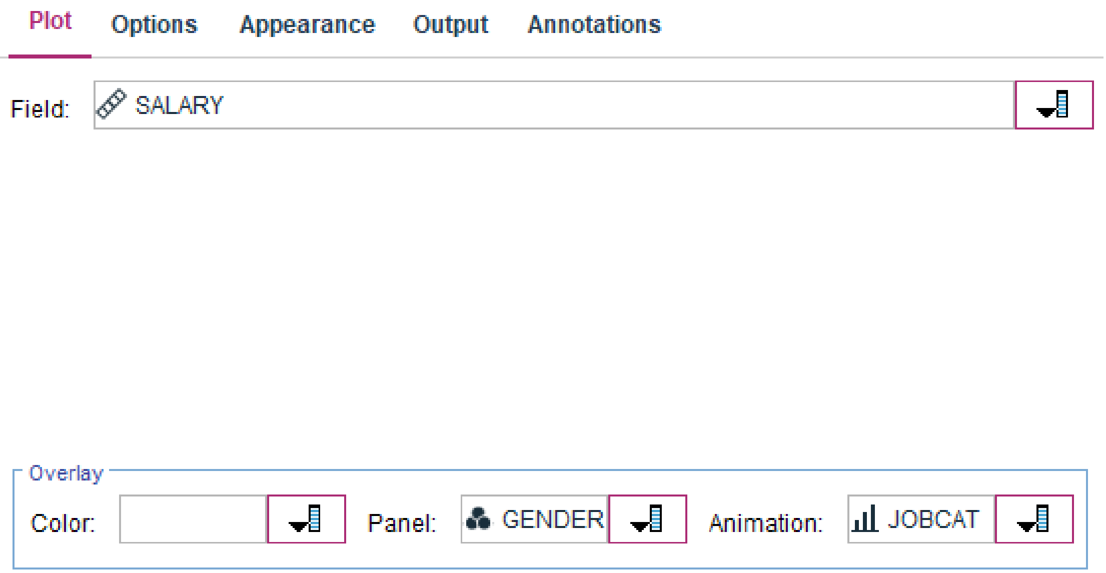


Output


100,000
80,000
60,000
40,000
20,000
0
1988-01-01
1990-01-01
1992-01-01
1994-01-01
1996-01-01
1998-01-01
date
men
women
jewel
1989.0
1998.0
1989.0
1990.0
1991.0
1992.0
1993.0
1994.0
1995.0
1996.0
1997.0
1998.0

Histogram node

1. Add **Statistics file** from **Source** palette;
2. Import *“employee_data.sav”*;
3. Add **Histogram** node from **Graphs** palette;
4. Add *“SALARY”* to **Field**;
5. Add *“JOBCAT”* to **Animation** and *“Gender”* to **Panel**;
6. Click **Run**.

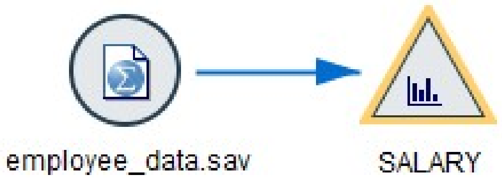


Field: SALARY

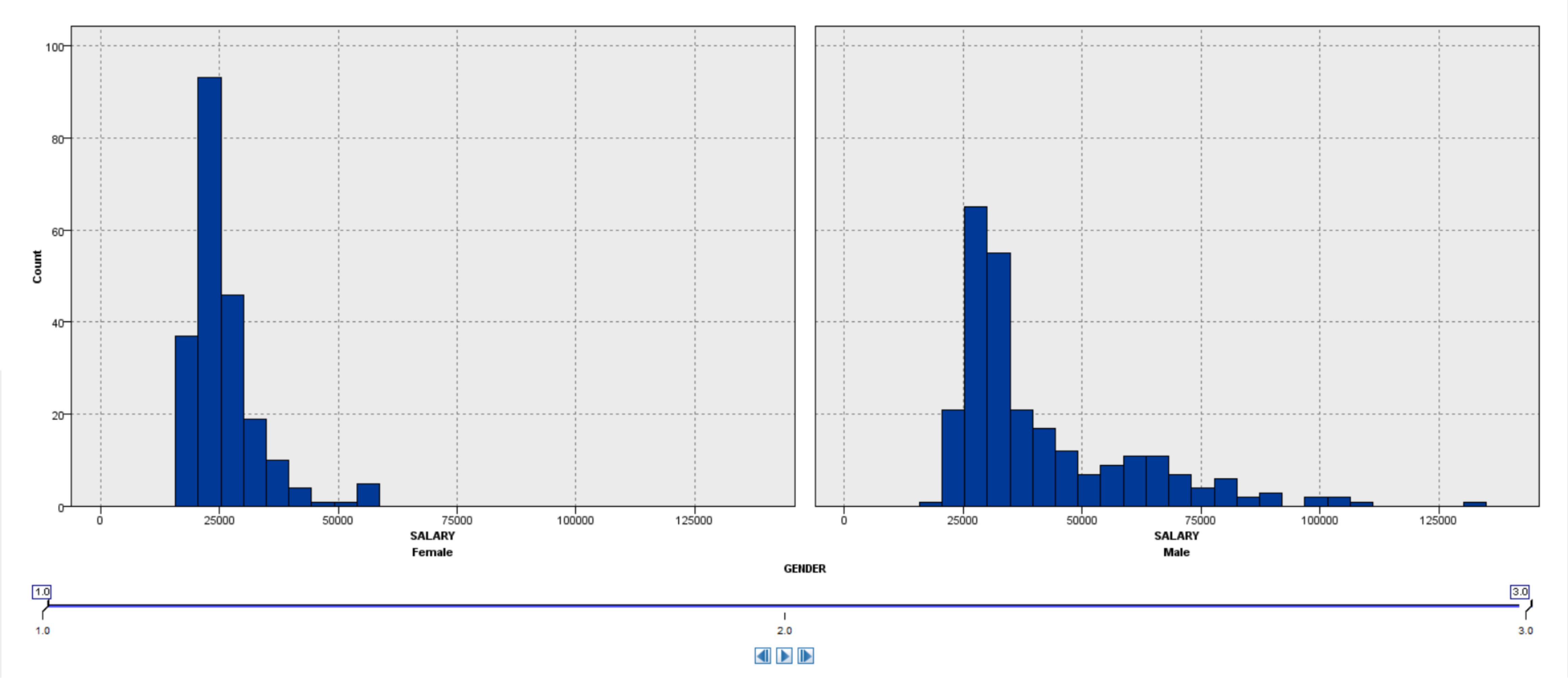
Color:

Panel: GENDER

Animation: JOBCAT

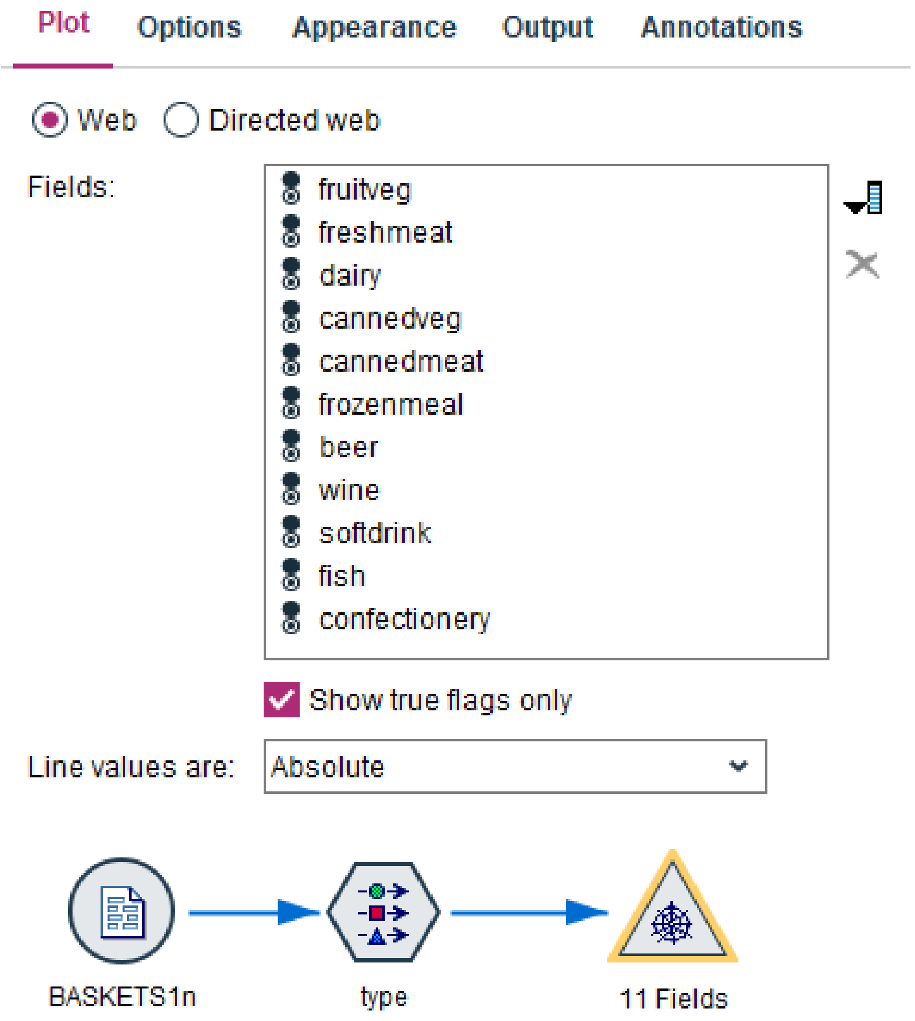


Output


Count
100
80
60
40
20
0
0
25000
50000
75000
100000
125000
SALARY
Female
SALARY
Male
GENDER
1.0
2.0
3.0

Web node

1. Add **Var. file** from **Source** palette;
2. Import *"BASKETS1n.txt"*;
3. Add **Type** node from **Field Ops** palette;
4. Double click on **Type** node and click **Read Values**;
5. Add **Web** node from **Graphs** palette;
6. Add *"fruitveg,...,confectionery"* to **Fields**;
7. Select **"Show true flags only"**;
8. Click **Run**.



● Web ○ Directed web

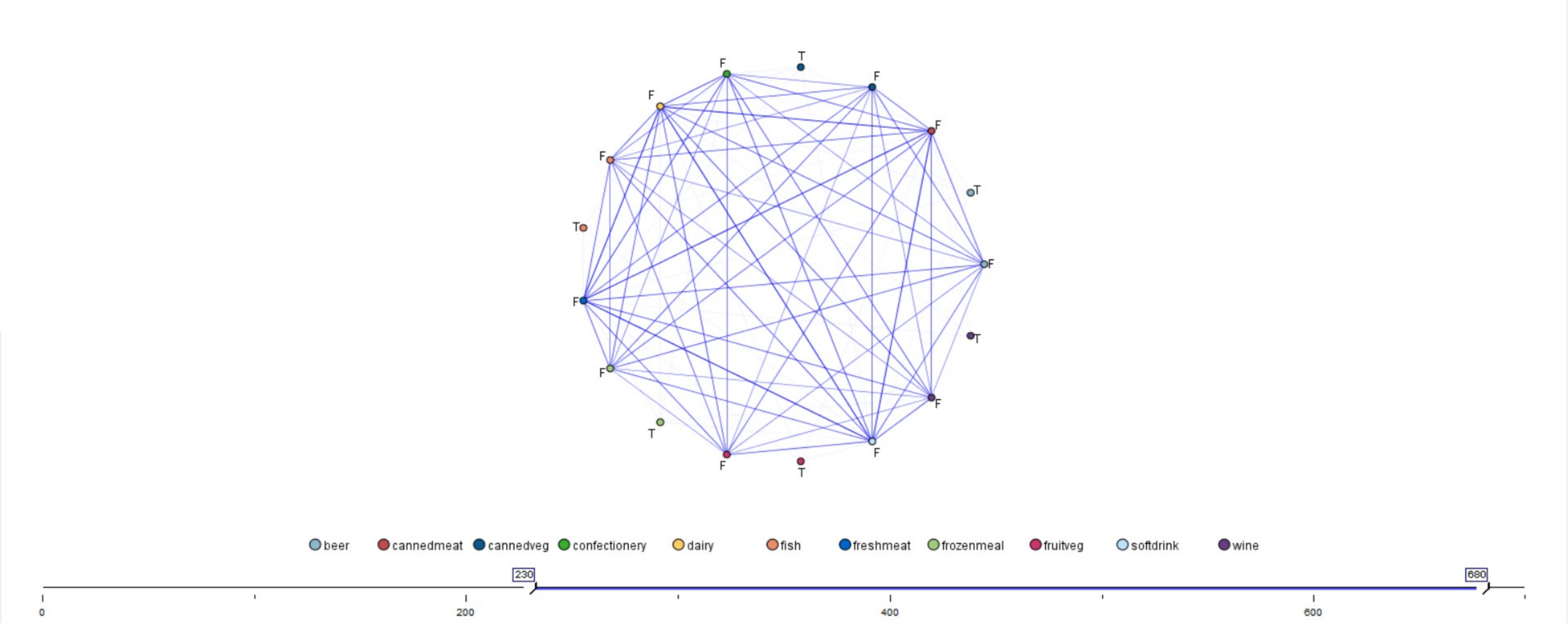
Fields:

- fruitveg
- freshmeat
- dairy
- cannedveg
- cannedmeat
- frozenmeal
- beer
- wine
- softdrink
- fish
- confectionery

☑ Show true flags only

Line values are: Absolute

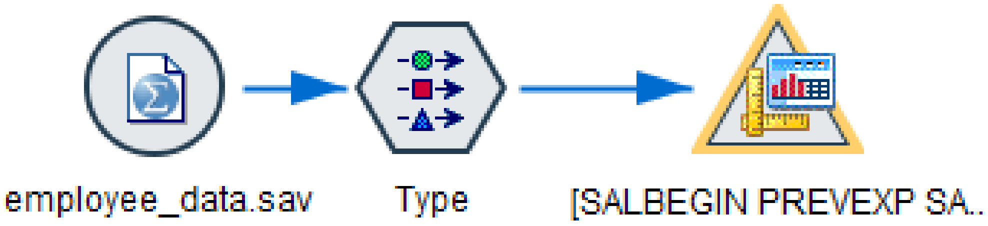
Output



T
F
F
F
F
F
T
To
F
F
F
T
F
F
T
F
T
F
beer
cannedmeat
cannedveg
confectionery
dairy
fish
freshmeat
frozenmeal
fruitveg
softdrink
wine
230
680
0
200
400
600

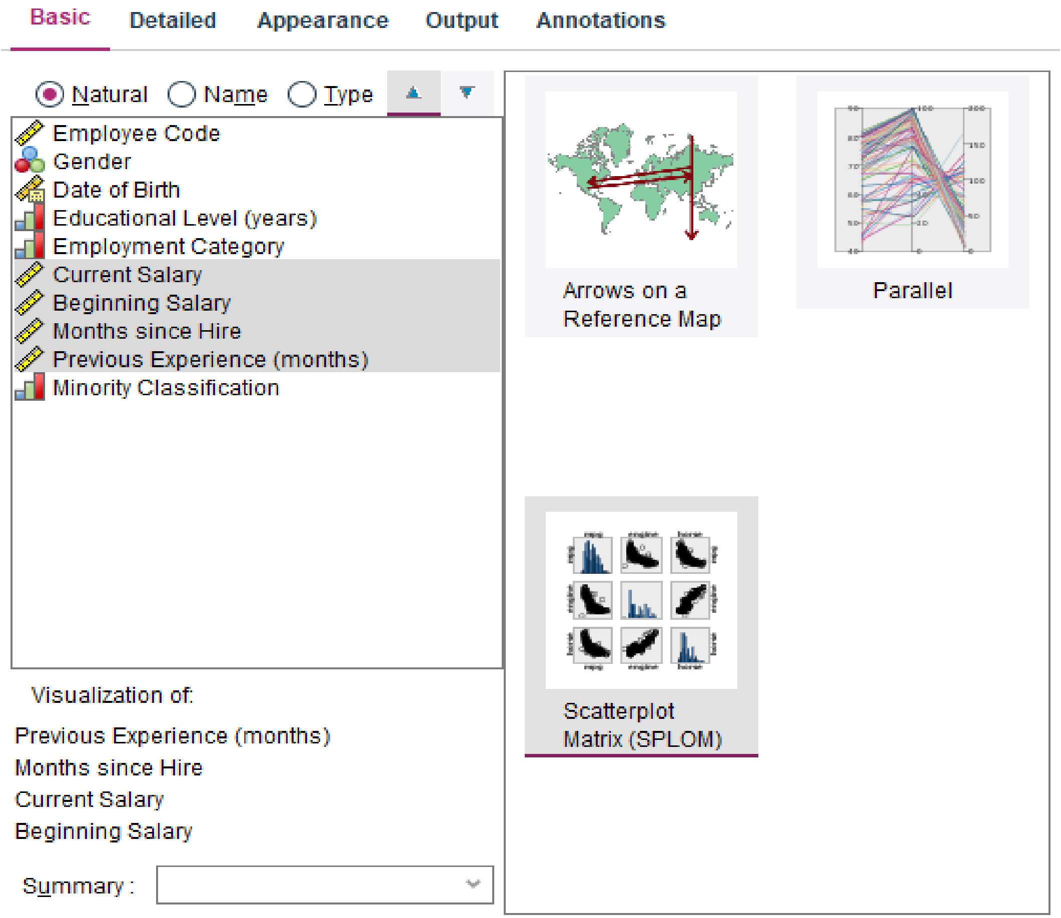
Scatterplot matrices with Graphboard

1. Add **Statistics** file from **Source** palette;
2. Import *“employee_data.sav”*;
3. Add **Type** node from **Field Ops** palette;
4. Click **Read Values**;
5. Add **Graphboard** node from **Graphs** palette;



Scatterplot matrices with Graphboard

6. Select *“Current Salary”, “Beginning Salary”, “Month since hire”, “Previous Experience”;*
7. Select **Scatterplot Matrix (SPLOM)**;
8. Click **Run**;



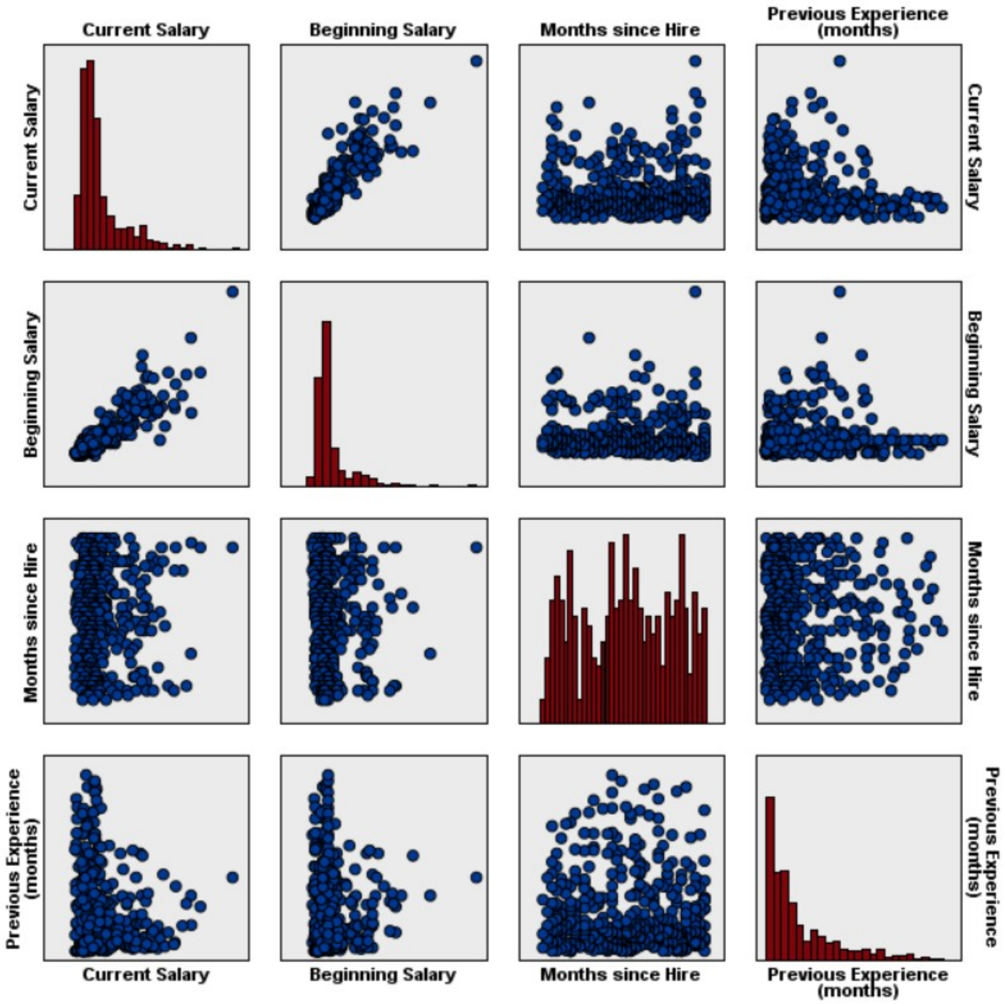
Employee Code
Gender
Date of Birth
Educational Level (years)
Employment Category
Current Salary
Beginning Salary
Months since Hire
Previous Experience (months)
Minority Classification

Visualization of:

Previous Experience (months)
Months since Hire
Current Salary
Beginning Salary

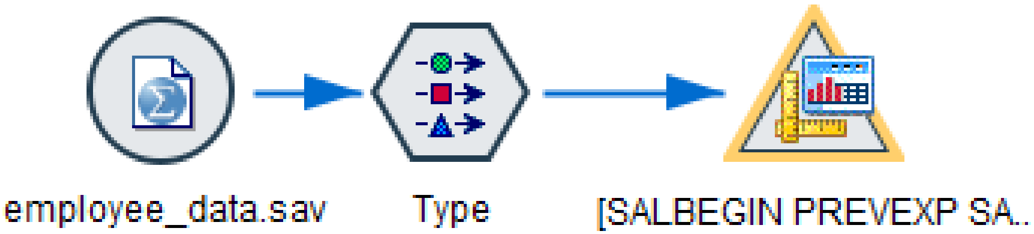
Summary :

Output


Current Salary
Beginning Salary
Months since Hire
Previous Experience (months)
Current Salary
Beginning Salary
Months since Hire
Previous Experience (months)
Current Salary
Beginning Salary
Months since Hire
Previous Experience (months)
Current Salary
Beginning Salary
Months since Hire
Previous Experience (months)

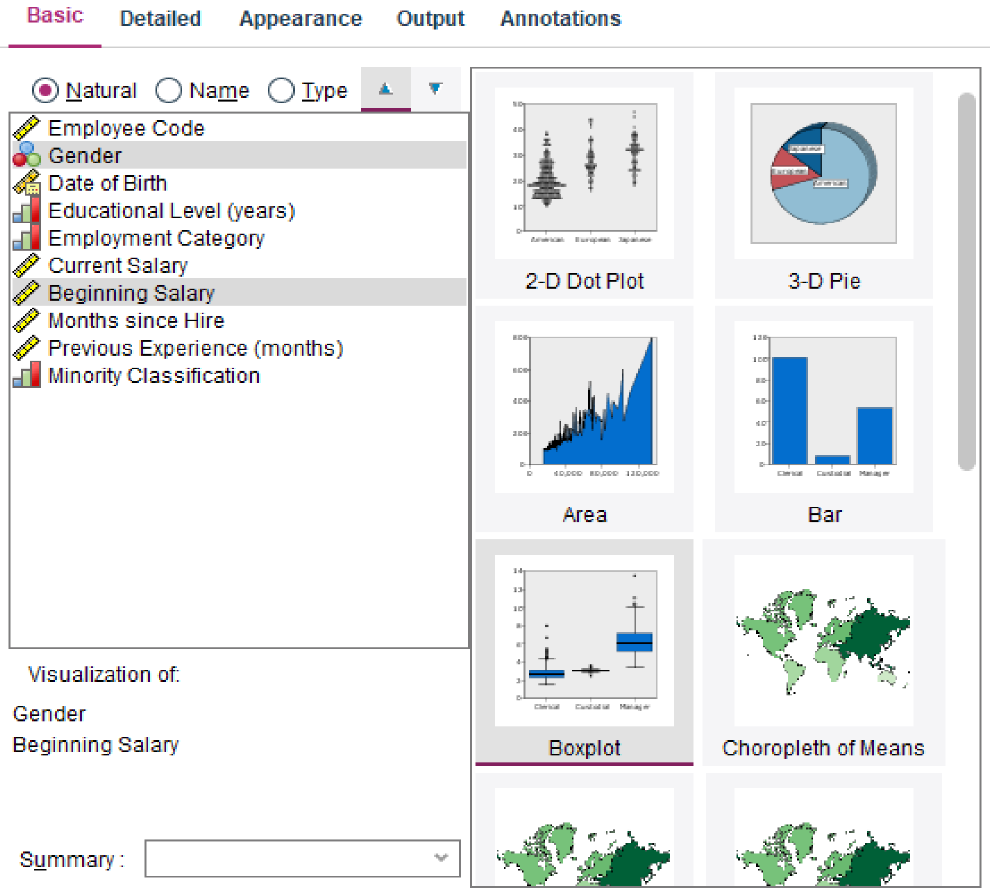
Boxplot with Graphboard

1. Add **Statistics** file from **Source** palette;
2. Import *“employee_data.sav”*;
3. Add **Type** node from **Field Ops** palette;
4. Click **Read Values**;



Boxplot with Graphboard

5. Add **Graphboard** node from **Graphs** palette;
6. Select *“Current Salary”*, *“Gender”*;
7. Select *“Boxplot”*;
8. Click **Run**

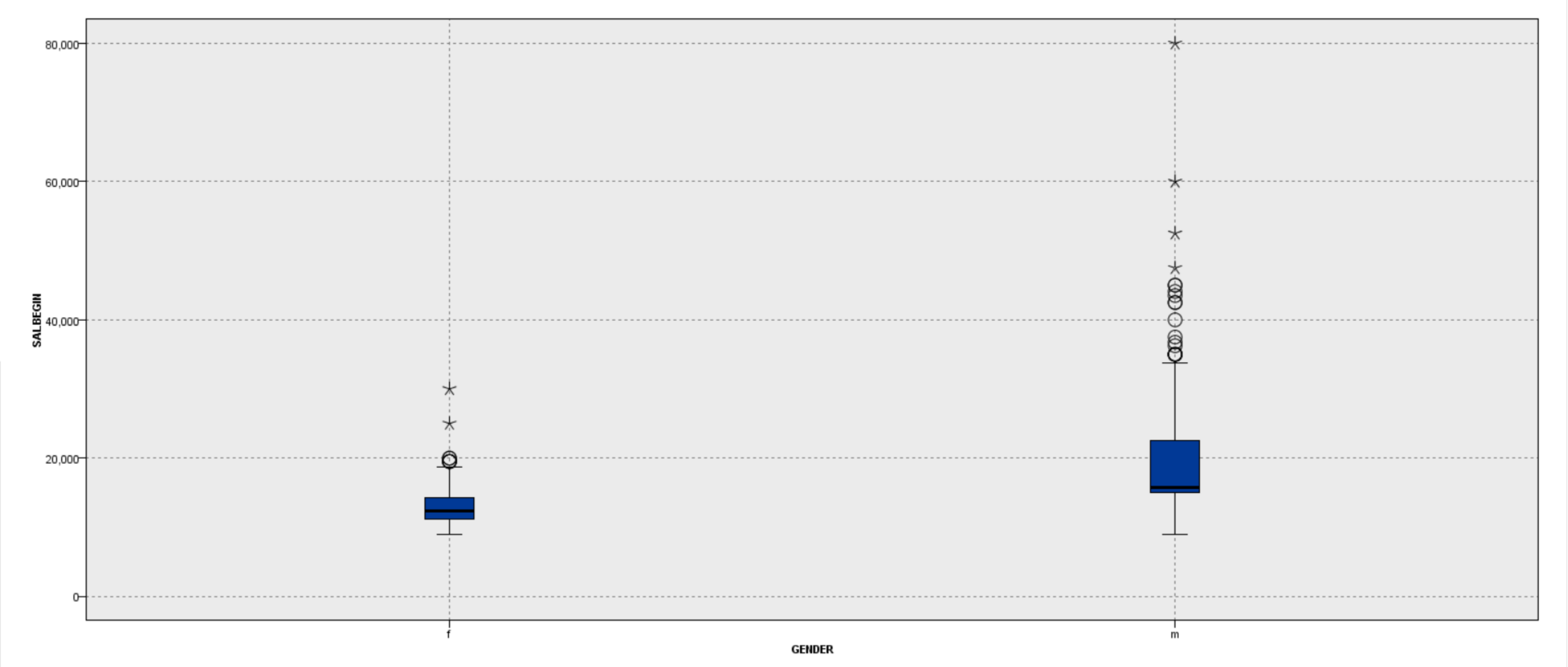


Employee Code
Gender
Date of Birth
Educational Level (years)
Employment Category
Current Salary
Beginning Salary
Months since Hire
Previous Experience (months)
Minority Classification

Visualization of:

Gender
Beginning Salary

Output



Thank you!

Data Science Academy