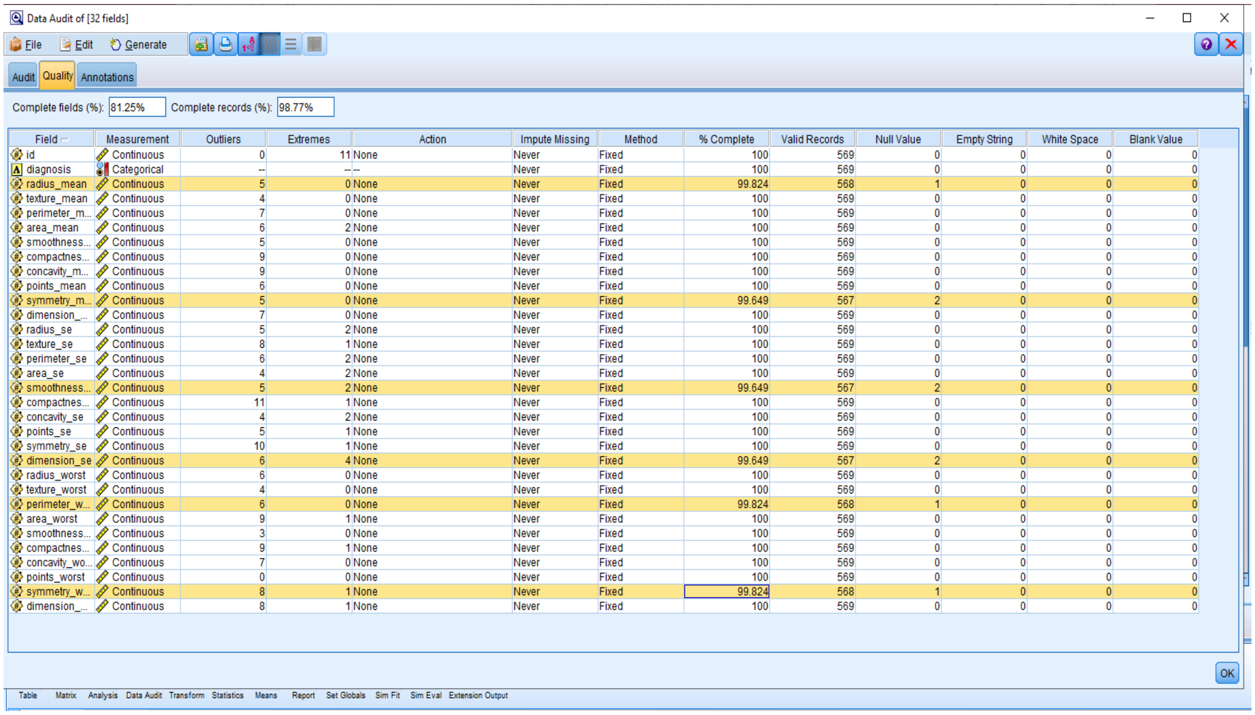
1. Explore whether there are missing values for any of the variables. Show how you did so in SPSS Modeler

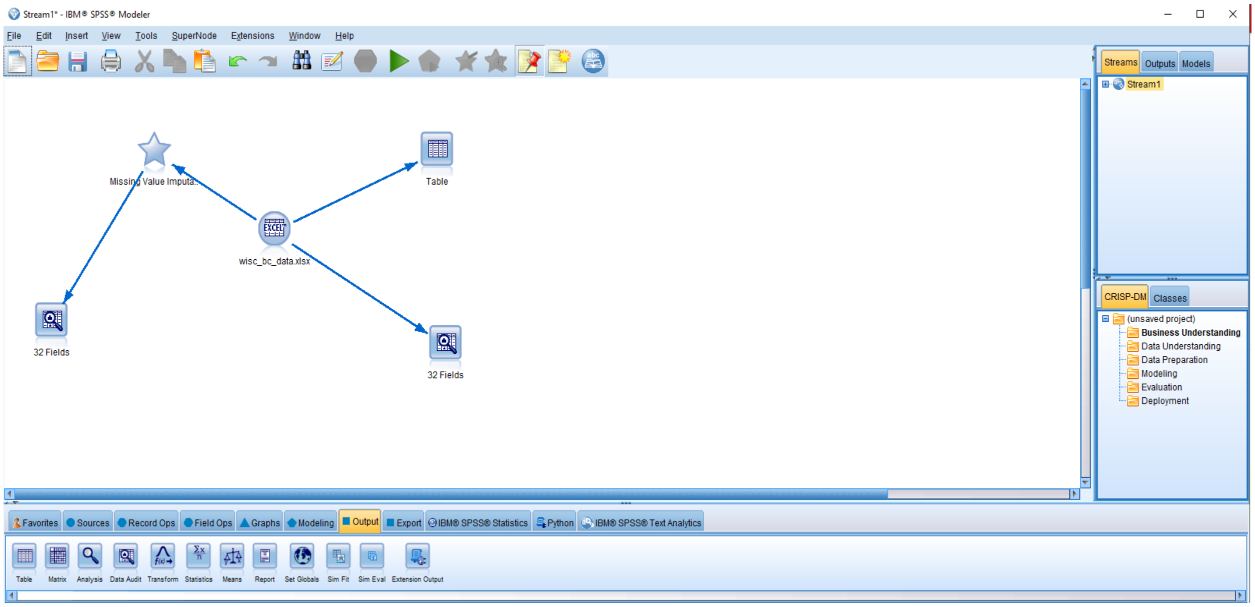
Below is the snapshot for all the missing variables from the given data set, we have used the features of Data Audit output table to get the quantity of missing variables.



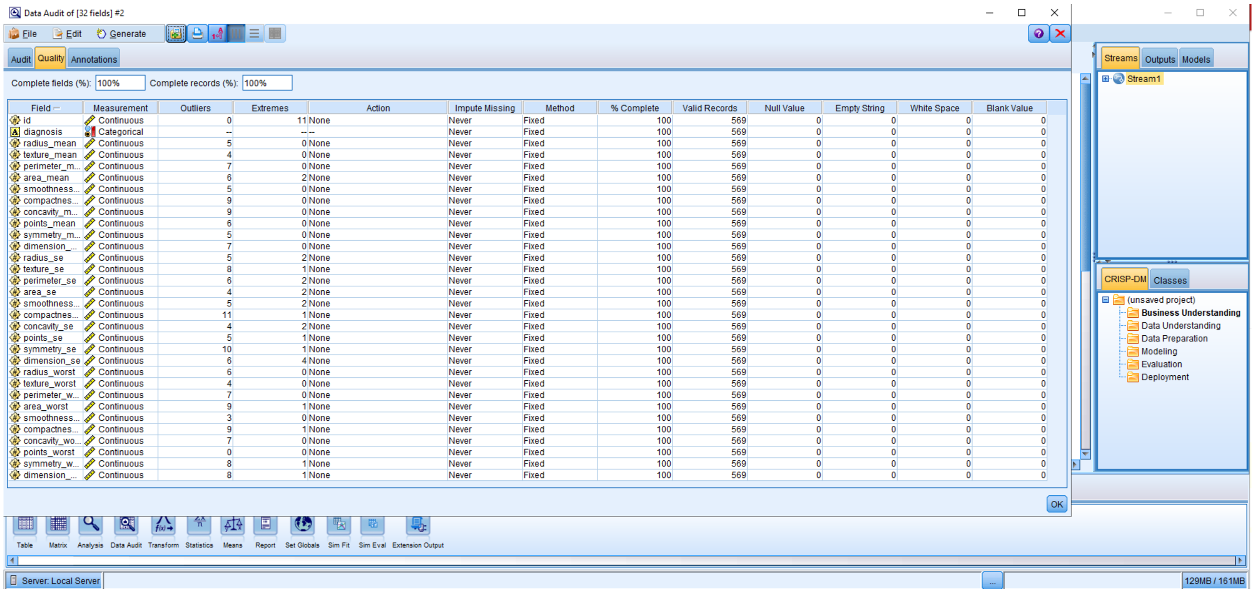
Complete fields (%): 81.25% Complete records (%): 98.77%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
id	Continuous	0	11	None	Never	Fixed	100	569	0	0	0	0
diagnosis	Categorical	--	--	--	Never	Fixed	100	569	0	0	0	0
radius_mean	Continuous	5	0	None	Never	Fixed	99.824	568	1	0	0	0
texture_mean	Continuous	4	0	None	Never	Fixed	100	569	0	0	0	0
perimeter_m...	Continuous	7	0	None	Never	Fixed	100	569	0	0	0	0
area_mean	Continuous	6	2	None	Never	Fixed	100	569	0	0	0	0
smoothness...	Continuous	5	0	None	Never	Fixed	100	569	0	0	0	0
compactnes...	Continuous	9	0	None	Never	Fixed	100	569	0	0	0	0
concavity_m...	Continuous	9	0	None	Never	Fixed	100	569	0	0	0	0
points_mean	Continuous	6	0	None	Never	Fixed	100	569	0	0	0	0
symmetry_m...	Continuous	5	0	None	Never	Fixed	99.649	567	2	0	0	0
dimension_...	Continuous	7	0	None	Never	Fixed	100	569	0	0	0	0
radius_se	Continuous	5	2	None	Never	Fixed	100	569	0	0	0	0
texture_se	Continuous	8	1	None	Never	Fixed	100	569	0	0	0	0
perimeter_se	Continuous	6	2	None	Never	Fixed	100	569	0	0	0	0
area_se	Continuous	4	2	None	Never	Fixed	100	569	0	0	0	0
smoothness...	Continuous	5	2	None	Never	Fixed	99.649	567	2	0	0	0
compactnes...	Continuous	11	1	None	Never	Fixed	100	569	0	0	0	0
concavity_se	Continuous	4	2	None	Never	Fixed	100	569	0	0	0	0
points_se	Continuous	5	1	None	Never	Fixed	100	569	0	0	0	0
symmetry_se	Continuous	10	1	None	Never	Fixed	100	569	0	0	0	0
dimension_se	Continuous	6	4	None	Never	Fixed	99.649	567	2	0	0	0
radius_worst	Continuous	6	0	None	Never	Fixed	100	569	0	0	0	0
texture_worst	Continuous	4	0	None	Never	Fixed	100	569	0	0	0	0
perimeter_w...	Continuous	6	0	None	Never	Fixed	99.824	568	1	0	0	0
area_worst	Continuous	9	1	None	Never	Fixed	100	569	0	0	0	0
smoothness...	Continuous	3	0	None	Never	Fixed	100	569	0	0	0	0
compactnes...	Continuous	9	1	None	Never	Fixed	100	569	0	0	0	0
concavity_wo...	Continuous	7	0	None	Never	Fixed	100	569	0	0	0	0
points_worst	Continuous	0	0	None	Never	Fixed	100	569	0	0	0	0
symmetry_w...	Continuous	8	1	None	Never	Fixed	99.824	568	1	0	0	0
dimension_...	Continuous	8	1	None	Never	Fixed	100	569	0	0	0	0

Once the missing variables has been handled, then we have generated the missing value node to address these variables from given data set.



Below is the snapshot of data after handling the missing variables, complete detail of variable is available in submitted stream.

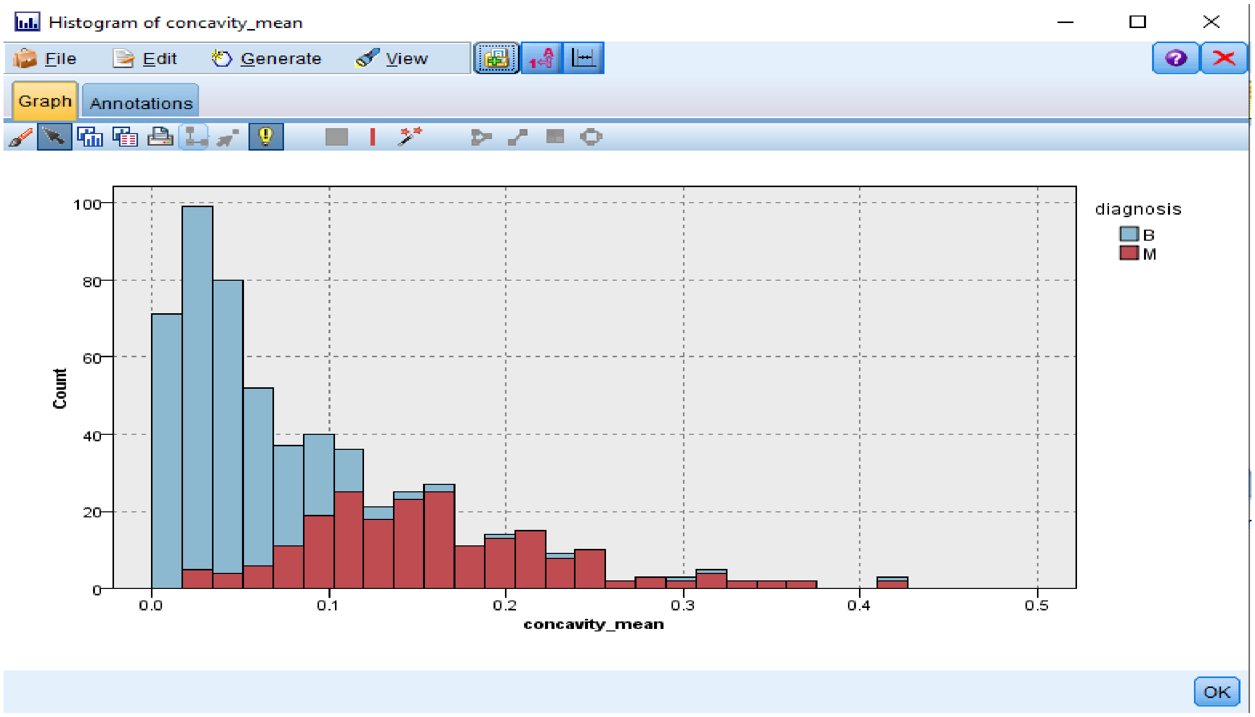


Complete fields (%): 100% Complete records (%): 100%

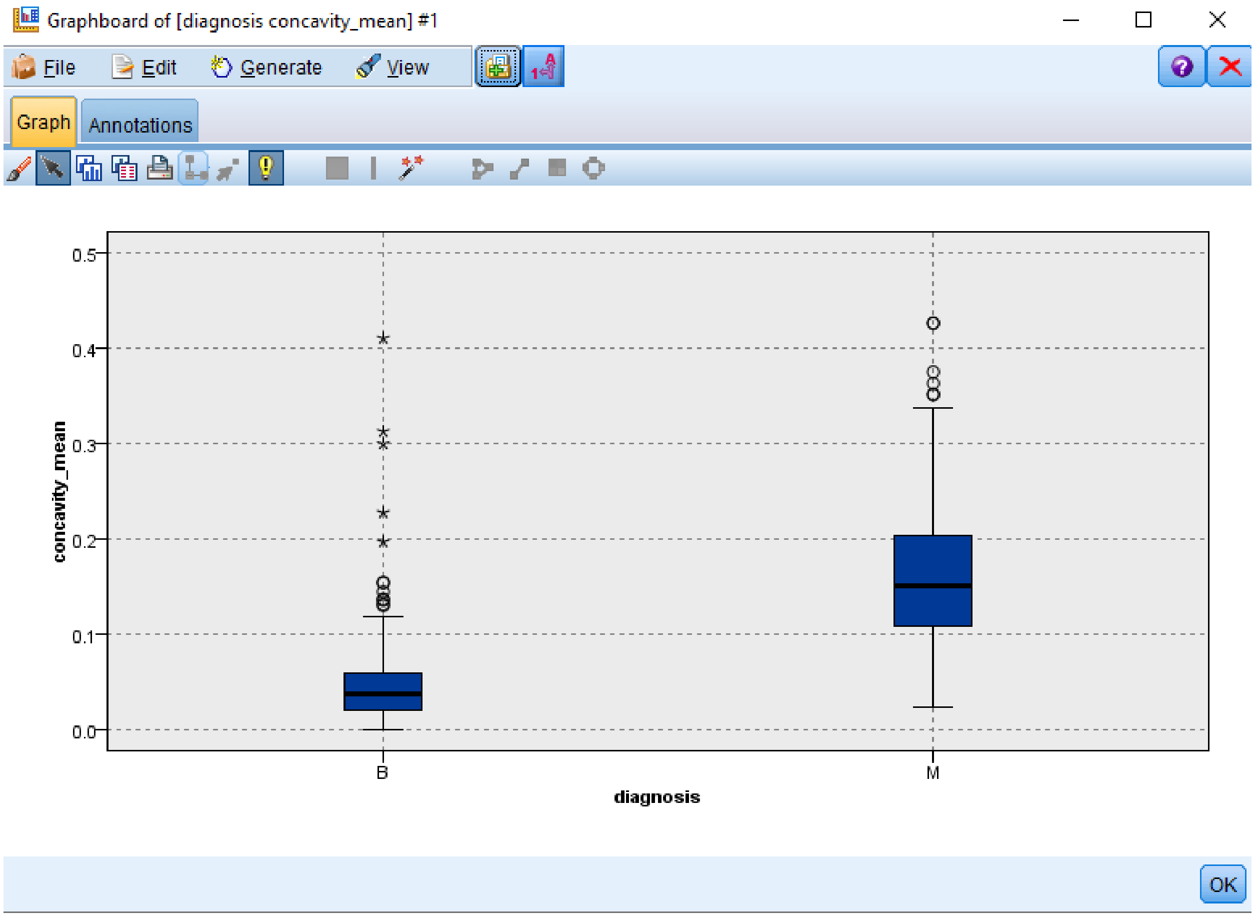
Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
id	Continuous	0	11	None	Never	Fixed	100	569	0	0	0	0
diagnosis	Categorical	--	--	--	Never	Fixed	100	569	0	0	0	0
radius_mean	Continuous	5	0	None	Never	Fixed	100	569	0	0	0	0
texture_mean	Continuous	4	0	None	Never	Fixed	100	569	0	0	0	0
perimeter_m...	Continuous	7	0	None	Never	Fixed	100	569	0	0	0	0
area_mean	Continuous	6	2	None	Never	Fixed	100	569	0	0	0	0
smoothness...	Continuous	5	0	None	Never	Fixed	100	569	0	0	0	0
compactnes...	Continuous	9	0	None	Never	Fixed	100	569	0	0	0	0
concavity_m...	Continuous	9	0	None	Never	Fixed	100	569	0	0	0	0
points_mean	Continuous	6	0	None	Never	Fixed	100	569	0	0	0	0
symmetry_m...	Continuous	5	0	None	Never	Fixed	100	569	0	0	0	0
dimension_...	Continuous	7	0	None	Never	Fixed	100	569	0	0	0	0
radius_se	Continuous	5	2	None	Never	Fixed	100	569	0	0	0	0
texture_se	Continuous	8	1	None	Never	Fixed	100	569	0	0	0	0
perimeter_se	Continuous	6	2	None	Never	Fixed	100	569	0	0	0	0
area_se	Continuous	4	2	None	Never	Fixed	100	569	0	0	0	0
smoothness...	Continuous	5	2	None	Never	Fixed	100	569	0	0	0	0
compactnes...	Continuous	11	1	None	Never	Fixed	100	569	0	0	0	0
concavity_se	Continuous	4	2	None	Never	Fixed	100	569	0	0	0	0
points_se	Continuous	5	1	None	Never	Fixed	100	569	0	0	0	0
symmetry_se	Continuous	10	1	None	Never	Fixed	100	569	0	0	0	0
dimension_se	Continuous	6	4	None	Never	Fixed	100	569	0	0	0	0
radius_worst	Continuous	6	0	None	Never	Fixed	100	569	0	0	0	0
texture_worst	Continuous	4	0	None	Never	Fixed	100	569	0	0	0	0
perimeter_w...	Continuous	7	0	None	Never	Fixed	100	569	0	0	0	0
area_worst	Continuous	9	1	None	Never	Fixed	100	569	0	0	0	0
smoothness...	Continuous	3	0	None	Never	Fixed	100	569	0	0	0	0
compactnes...	Continuous	9	1	None	Never	Fixed	100	569	0	0	0	0
concavity_wo...	Continuous	7	0	None	Never	Fixed	100	569	0	0	0	0
points_worst	Continuous	0	0	None	Never	Fixed	100	569	0	0	0	0
symmetry_w...	Continuous	8	1	None	Never	Fixed	100	569	0	0	0	0
dimension_...	Continuous	8	1	None	Never	Fixed	100	569	0	0	0	0

2. Use a graph to determine visually whether there are any outliers in the average concavity field (concavity_mean) that measures the severity of concave portions of the contour. What kind of graph would provide better visualization for this task.

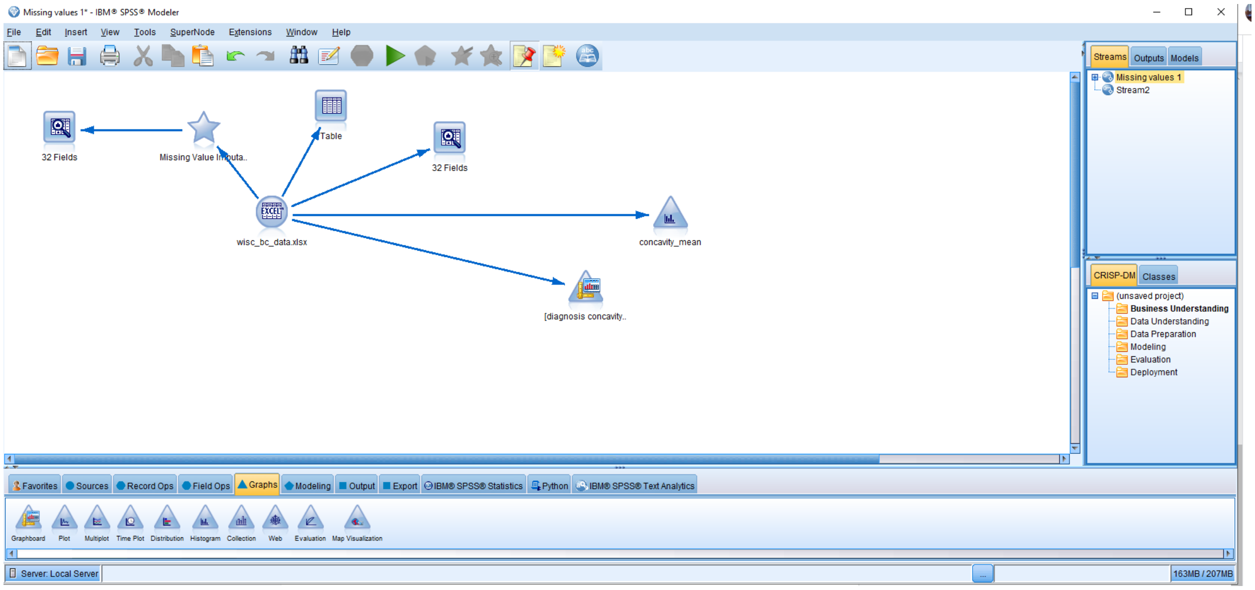
Below is visual representation of the average concavity attribute (concavity_mean), from the below histogram graph we can say the there is no outliers available, we will also check with box plot graph.



Below is the box plot graph representation of the average concavity attribute (concavity_mean).



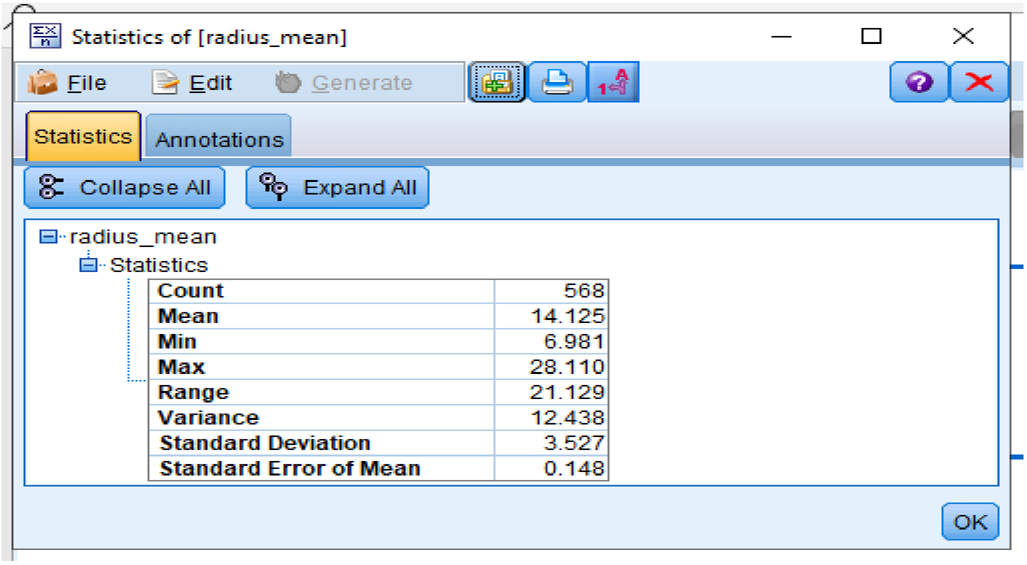
From a above box plot, we can confirm that there are no outliers available.



The above is the snapshot of stream till this task.

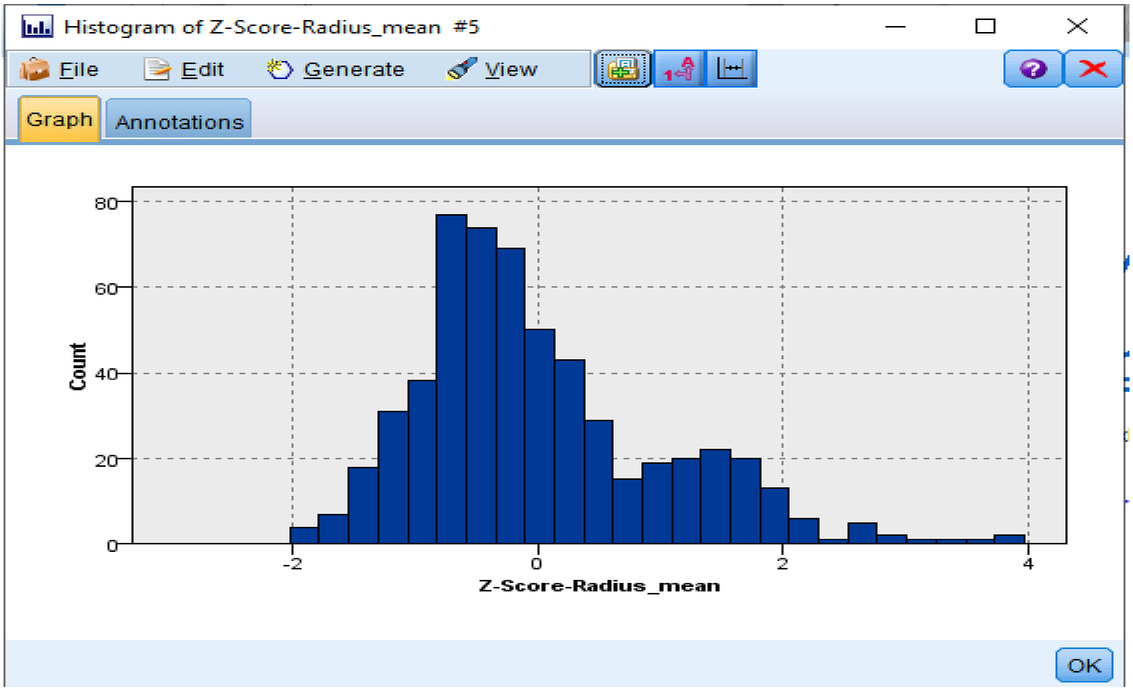
3. Transform the radius_mean attribute using Z-score standardization. Using a graph, describe the range of the standardized values. Hint: consider adding a Derive node to create the Z-score. You can find the required summary statistics (mean and std deviation) to calculate the Z-score using the statistics node.

In order to calculate the Z-score standardization, we first have to calculate the mean and standard deviation which can be shown below.

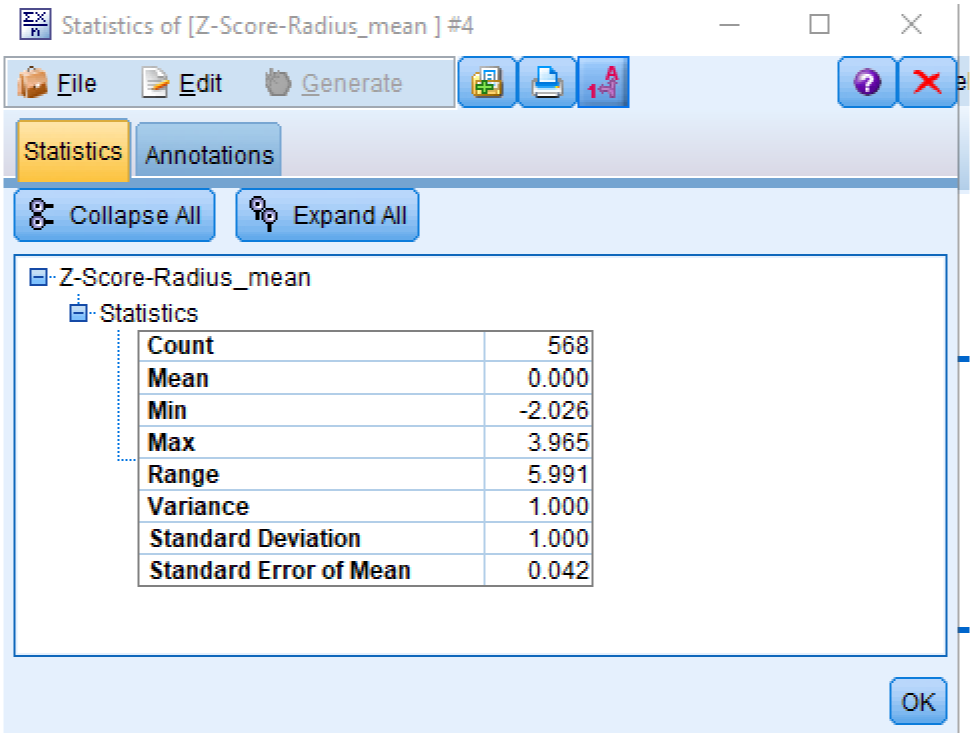


Count	568
Mean	14.125
Min	6.981
Max	28.110
Range	21.129
Variance	12.438
Standard Deviation	3.527
Standard Error of Mean	0.148

The above is mean and standard deviation for given data using statics, we will be using these values to calculate the asked z-score.



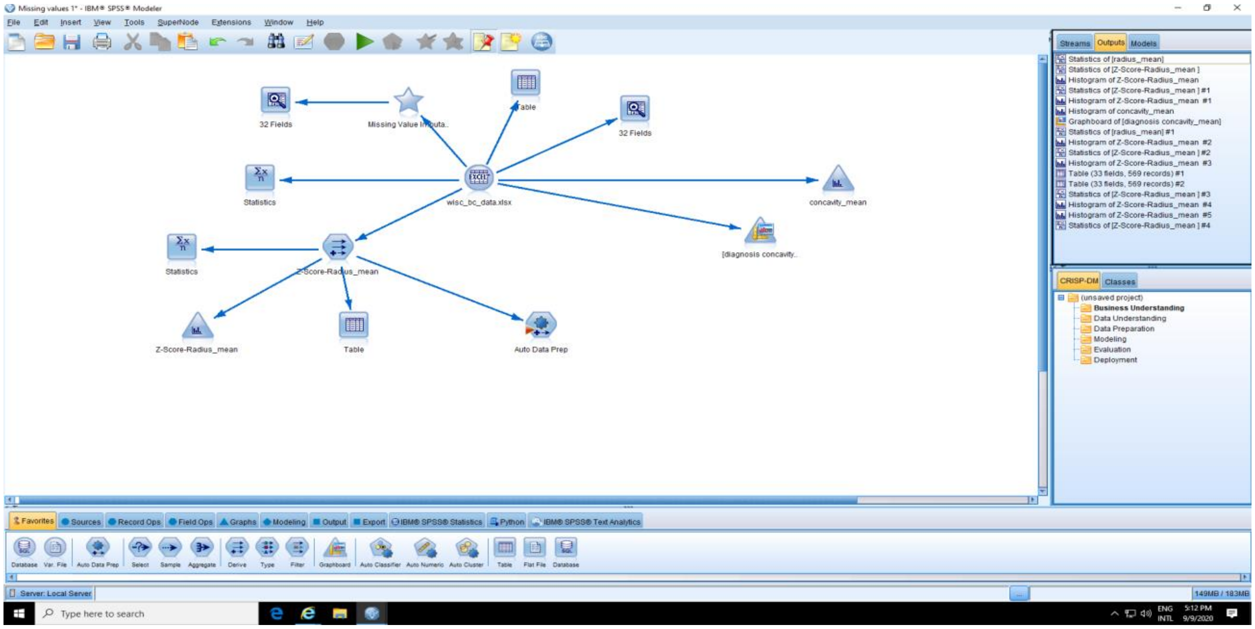
The above is the histogram graph representation of Z-score standardization, which mostly varies from -2 to 2.



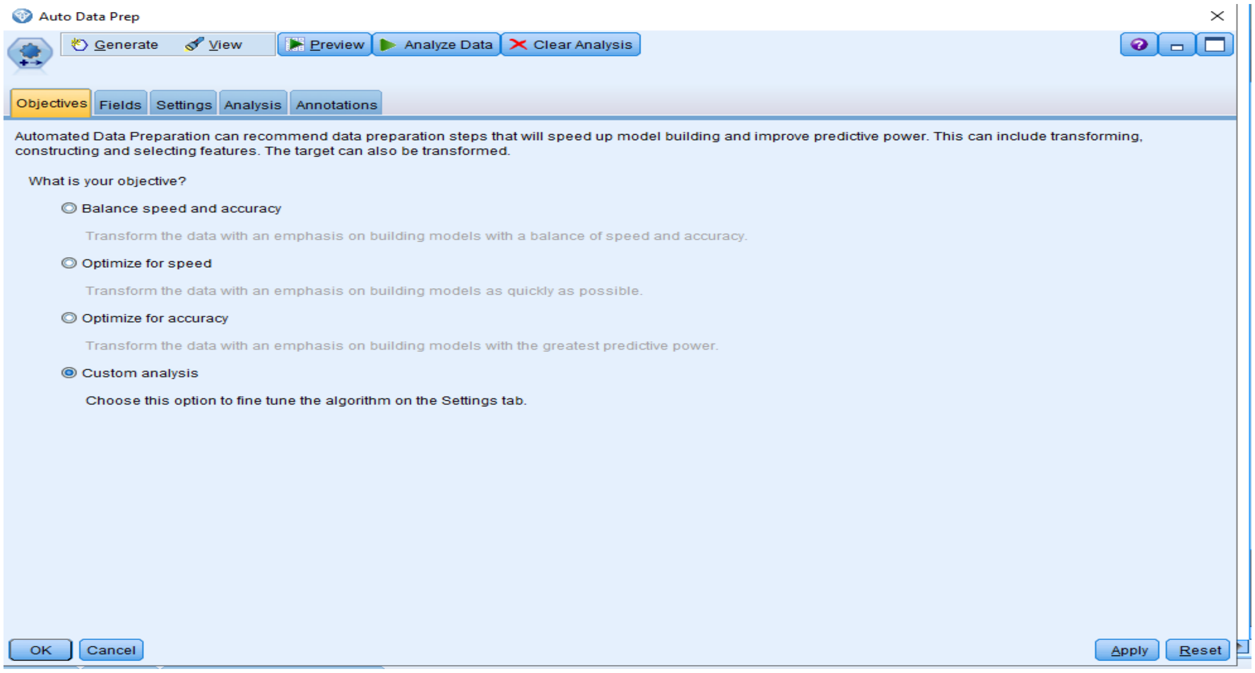
Count	568
Mean	0.000
Min	-2.026
Max	3.965
Range	5.991
Variance	1.000
Standard Deviation	1.000
Standard Error of Mean	0.042

We have transformed the radius_mean attribute using Z-score standardization (mean µ = 14.125) and standard deviation (σ) = 3.527. The Z-score ranges from about - 2.026 to 3.965 indicating the potential existence for several outliers.

4.How would you transform all 30 metrics to Z-scores without having to create derived nodes for each of them? Hint: Investigate the Auto Data Prep node. We did not cover this in class, but it should be very simple with a little experimentation.


Missing values 1* - IBM® SPSS® Modeler
File Edit Insert View Tools SuperNode Extensions Window Help
32 Fields
Missing Value Imputa...
Table
32 Fields
Statistics
wisc_bc_data.xlsx
concavity_mean
Statistics
Z-Score-Radius_mean
[diagnosis concavity...
Z-Score-Radius_mean
Table
Auto Data Prep
Streams Outputs Models
Statistics of [radius_mean]
Statistics of [Z-Score-Radius_mean]
Histogram of Z-Score-Radius_mean
Statistics of [Z-Score-Radius_mean] #1
Histogram of Z-Score-Radius_mean #1
Histogram of concavity_mean
Graphboard of [diagnosis concavity_mean]
Statistics of [radius_mean] #1
Histogram of Z-Score-Radius_mean #2
Statistics of [Z-Score-Radius_mean] #2
Histogram of Z-Score-Radius_mean #3
Table (33 fields, 569 records) #1
Table (33 fields, 569 records) #2
Statistics of [Z-Score-Radius_mean] #3
Histogram of Z-Score-Radius_mean #4
Histogram of Z-Score-Radius_mean #5
Statistics of [Z-Score-Radius_mean] #4
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM® SPSS® Statistics Python IBM® SPSS® Text Analytics
Database Var. File Auto Data Prep Select Sample Aggregate Derive Type Filter Graphboard Auto Classifier Auto Numeric Auto Cluster Table Flat File Database
Server: Local Server
149MB / 183MB
Type here to search
ENG INTL 5:12 PM 9/9/2020

In order to calculate the Z-score for all the 30 metrics, we first need to add and connect the Auto Data prep node, which is shown in the snap shot above.



Automated Data Preparation can recommend data preparation steps that will speed up model building and improve predictive power. This can include transforming, constructing and selecting features. The target can also be transformed.

What is your objective?

○ Balance speed and accuracy

Transform the data with an emphasis on building models with a balance of speed and accuracy.

○ Optimize for speed

Transform the data with an emphasis on building models as quickly as possible.

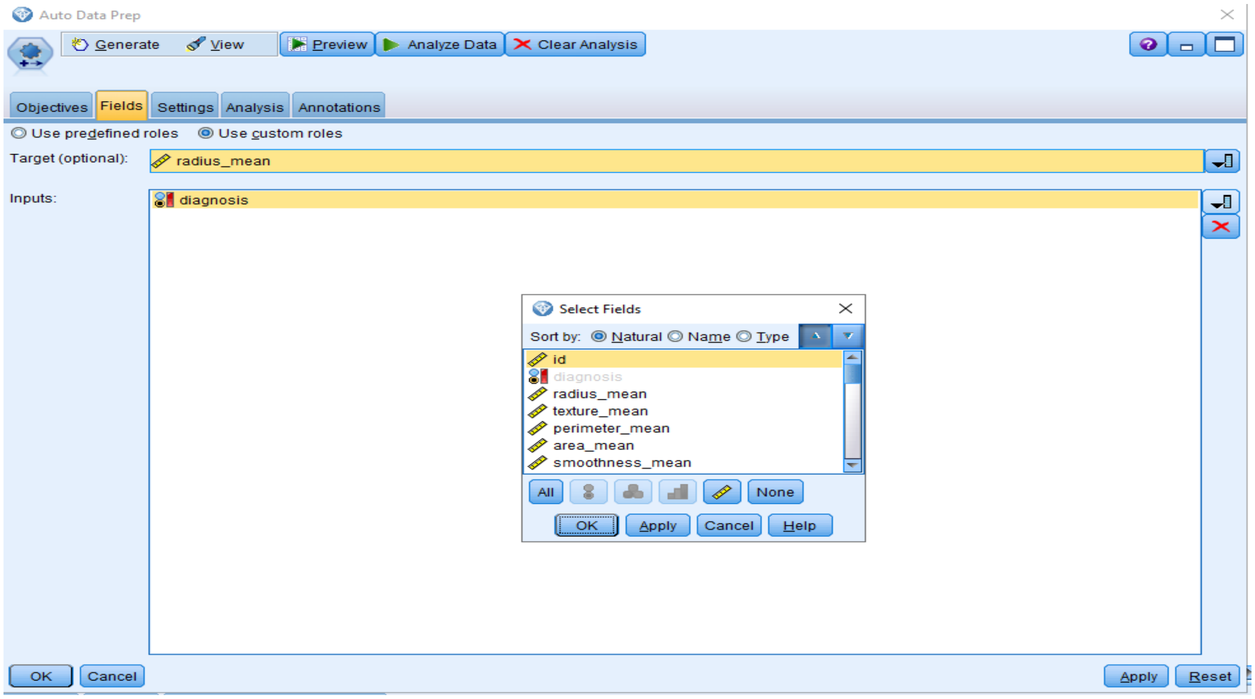
○ Optimize for accuracy

Transform the data with an emphasis on building models with the greatest predictive power.

● Custom analysis

Choose this option to fine tune the algorithm on the Settings tab.

From the edit option in Auto Data prep node, we have to select the custom analysis option, as shown above.



○ Use predefined roles ● Use custom roles

Target (optional): radius_mean

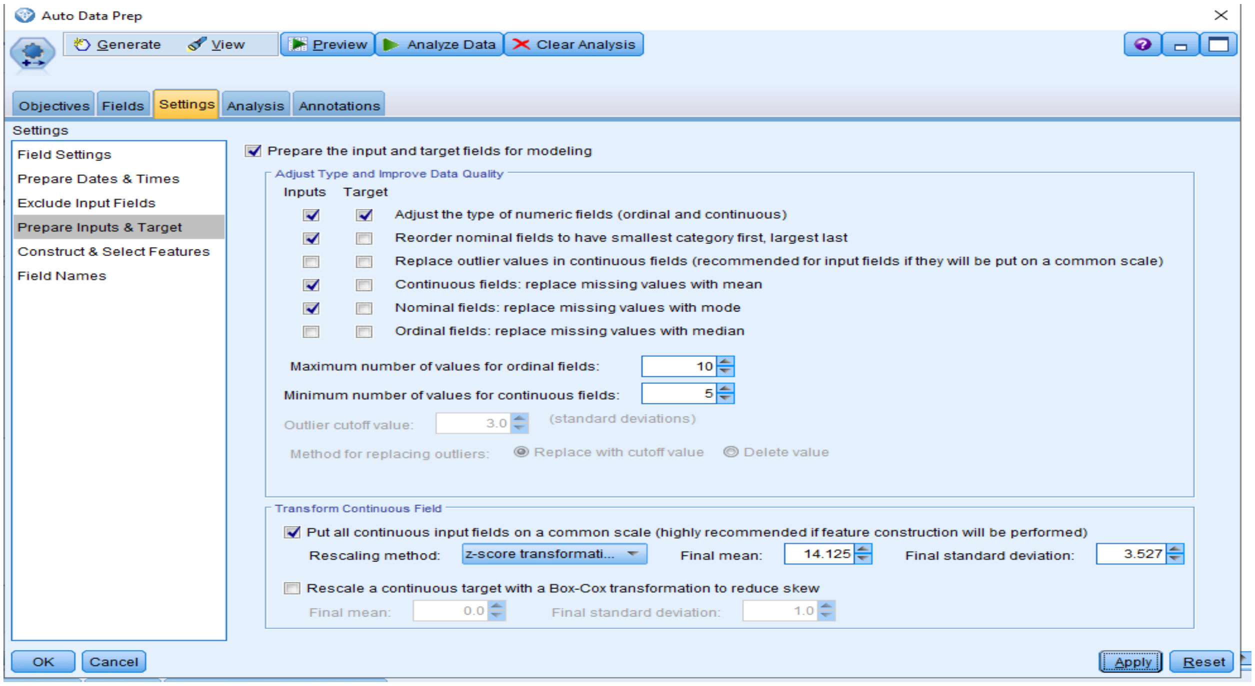
Inputs: diagnosis

Sort by: ● Natural ○ Name ○ Type

id
diagnosis
radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean

Under the fields option we can select the target and inputs field to calculate the required Z-score, we can select any attribute, to valid our calculation we have taken the radius_mean attribute whose Z-score we have already calculate in previous part.

Now under setting tab and in prepare Inputs and target section, we have to insert mean and standard deviation of the selected attribute, for this case we have, mean = 14.125 and standard deviation = 3.527.



☑ Prepare the input and target fields for modeling

Adjust Type and Improve Data Quality

Inputs	Target	
☑	☑	Adjust the type of numeric fields (ordinal and continuous)
☑	□	Reorder nominal fields to have smallest category first, largest last
□	□	Replace outlier values in continuous fields (recommended for input fields if they will be put on a common scale)
☑	□	Continuous fields: replace missing values with mean
☑	□	Nominal fields: replace missing values with mode
□	□	Ordinal fields: replace missing values with median

Maximum number of values for ordinal fields: 10

Minimum number of values for continuous fields: 5

Outlier cutoff value: 3.0 (standard deviations)

Method for replacing outliers: ● Replace with cutoff value ○ Delete value

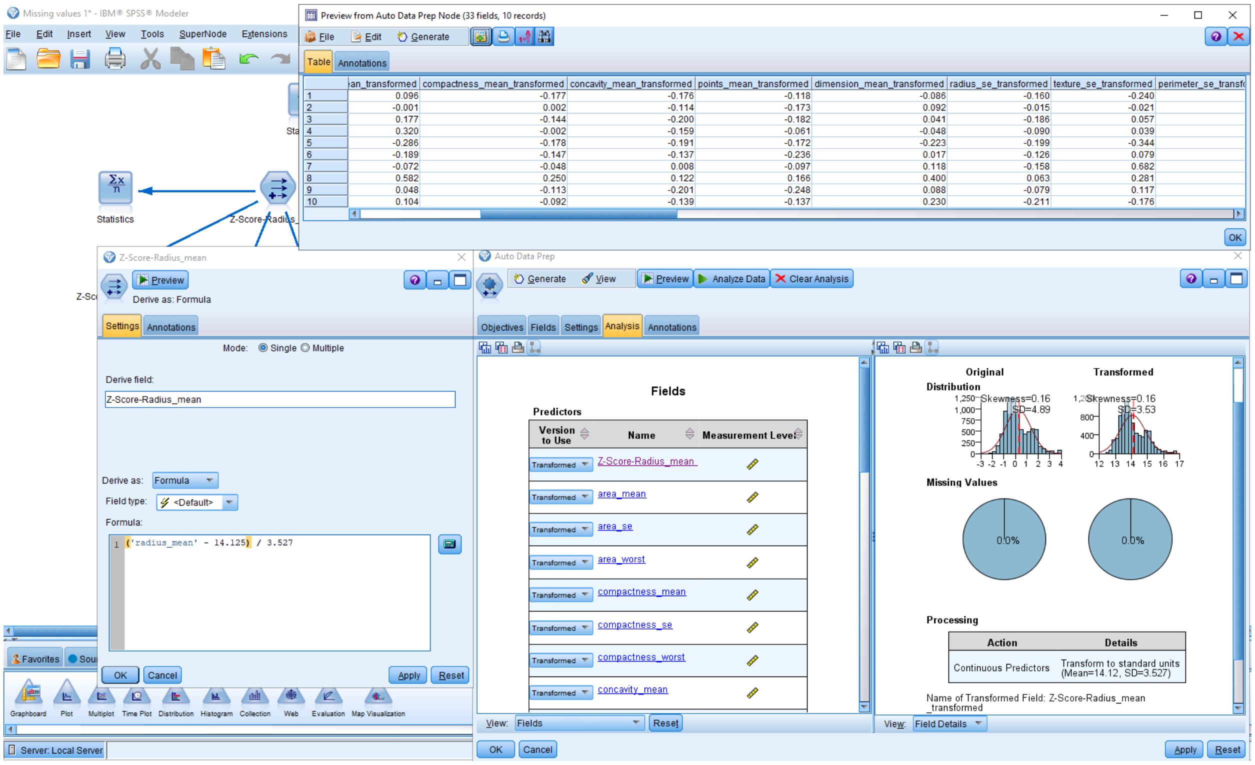
Transform Continuous Field

☑ Put all continuous input fields on a common scale (highly recommended if feature construction will be performed)

Rescaling method: z-score transformati... Final mean: 14.125 Final standard deviation: 3.527

□ Rescale a continuous target with a Box-Cox transformation to reduce skew

Final mean: 0.0 Final standard deviation: 1.0



Preview from Auto Data Prep Node (33 fields, 10 records)

	an_transformed	compactness_mean_transformed	concavity_mean_transformed	points_mean_transformed	dimension_mean_transformed	radius_se_transformed	texture_se_transformed	perimeter_se_transf
1	0.096	-0.177	-0.176	-0.118	-0.086	-0.160	-0.240	
2	-0.001	0.002	-0.114	-0.173	0.092	-0.015	-0.021	
3	0.177	-0.144	-0.200	-0.182	0.041	-0.186	0.057	
4	0.320	-0.002	-0.159	-0.061	-0.048	-0.090	0.039	
5	-0.286	-0.178	-0.191	-0.172	-0.223	-0.199	-0.344	
6	-0.189	-0.147	-0.137	-0.236	0.017	-0.126	0.079	
7	-0.072	-0.048	0.008	-0.097	0.118	-0.158	0.682	
8	0.582	0.250	0.122	0.166	0.400	0.063	0.281	
9	0.048	-0.113	-0.201	-0.248	0.088	-0.079	0.117	
10	0.104	-0.092	-0.139	-0.137	0.230	-0.211	-0.176	

Derive as: Formula

Mode: ● Single ○ Multiple

Derive field: Z-Score-Radius_mean

Derive as: Formula

Field type: <Default>

Formula:

```
('radius_mean' - 14.125) / 3.527
```

Fields

Predictors

Version to Use	Name	Measurement Level
Transformed	Z-Score-Radius_mean	📏
Transformed	area_mean	📏
Transformed	area_se	📏
Transformed	area_worst	📏
Transformed	compactness_mean	📏
Transformed	compactness_se	📏
Transformed	compactness_worst	📏
Transformed	concavity_mean	📏

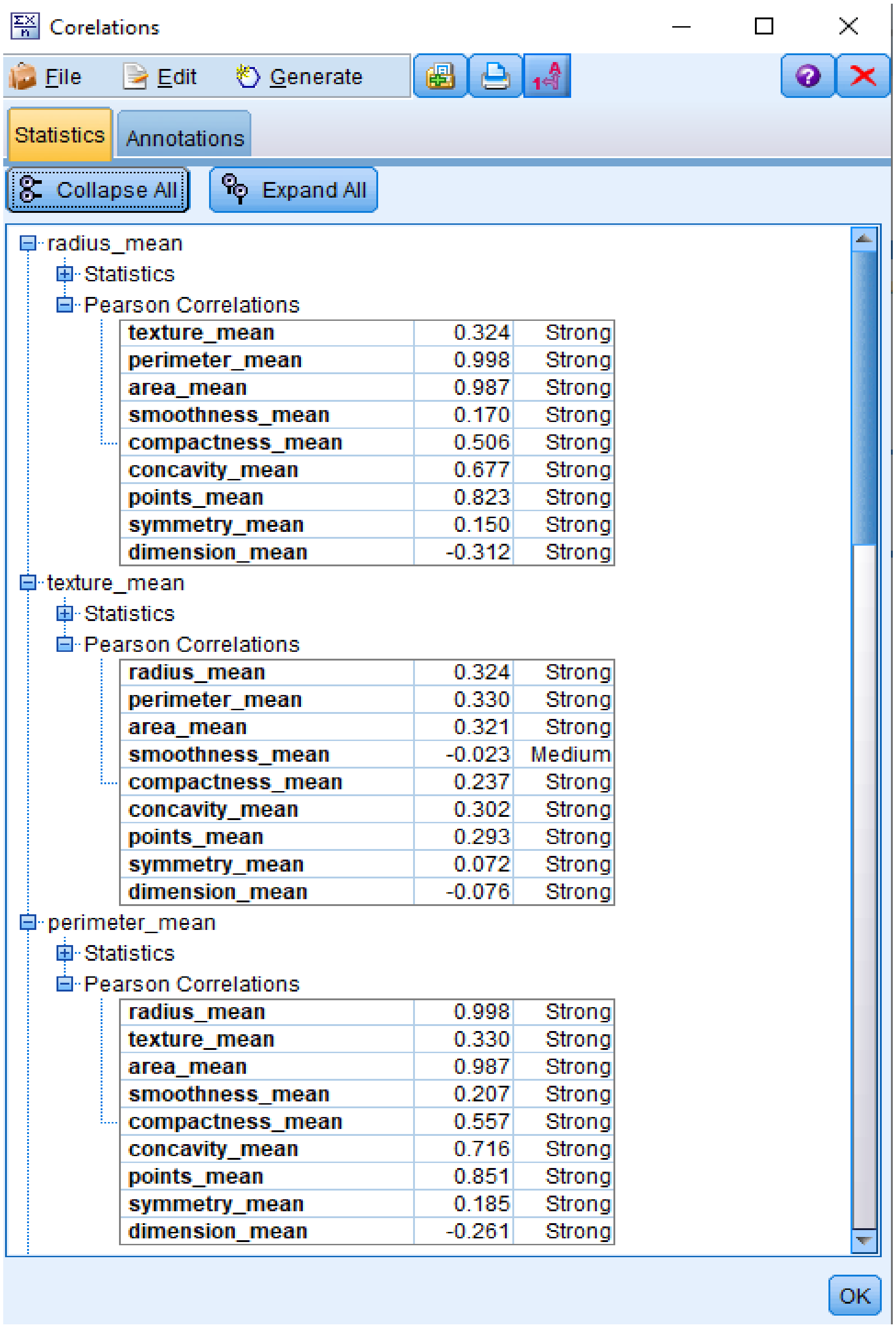
Processing

Action	Details
Continuous Predictors	Transform to standard units (Mean=14.12, SD=3.527)

Name of Transformed Field: Z-Score-Radius_mean _transformed

Once we click on Analyze Data from the above tab, we get the required results of selected attribute, this case the Z-score is varies in the same range which we have calculated earlier.

5. Investigate whether there are any correlated variables among the first 10 numeric fields (the ones with _mean prefix)



radius_mean

Statistics

Pearson Correlations

texture_mean	0.324	Strong
perimeter_mean	0.998	Strong
area_mean	0.987	Strong
smoothness_mean	0.170	Strong
compactness_mean	0.506	Strong
concavity_mean	0.677	Strong
points_mean	0.823	Strong
symmetry_mean	0.150	Strong
dimension_mean	-0.312	Strong

texture_mean

Statistics

Pearson Correlations

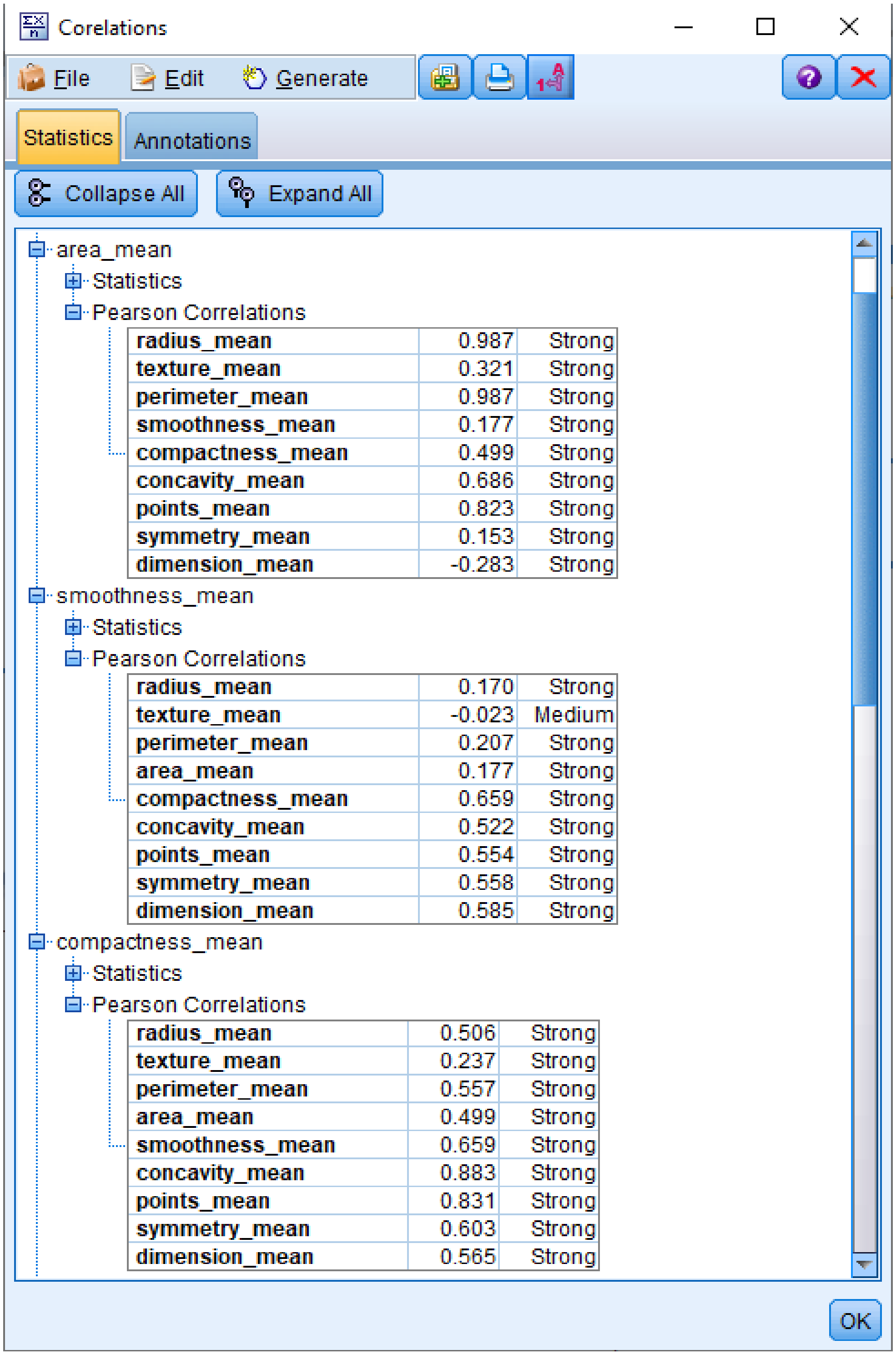
radius_mean	0.324	Strong
perimeter_mean	0.330	Strong
area_mean	0.321	Strong
smoothness_mean	-0.023	Medium
compactness_mean	0.237	Strong
concavity_mean	0.302	Strong
points_mean	0.293	Strong
symmetry_mean	0.072	Strong
dimension_mean	-0.076	Strong

perimeter_mean

Statistics

Pearson Correlations

radius_mean	0.998	Strong
texture_mean	0.330	Strong
area_mean	0.987	Strong
smoothness_mean	0.207	Strong
compactness_mean	0.557	Strong
concavity_mean	0.716	Strong
points_mean	0.851	Strong
symmetry_mean	0.185	Strong
dimension_mean	-0.261	Strong



area_mean

Statistics

Pearson Correlations

radius_mean	0.987	Strong
texture_mean	0.321	Strong
perimeter_mean	0.987	Strong
smoothness_mean	0.177	Strong
compactness_mean	0.499	Strong
concavity_mean	0.686	Strong
points_mean	0.823	Strong
symmetry_mean	0.153	Strong
dimension_mean	-0.283	Strong

smoothness_mean

Statistics

Pearson Correlations

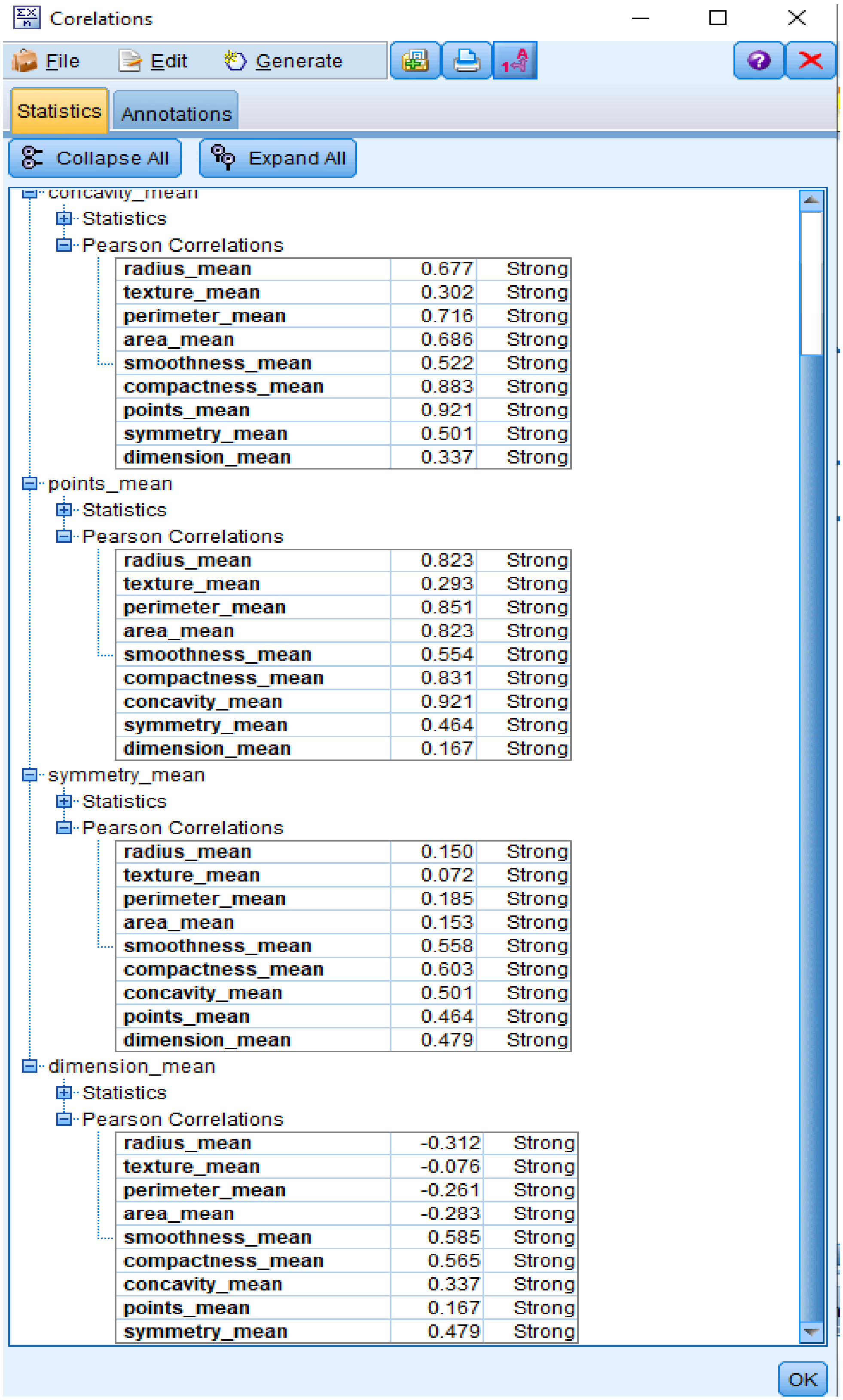
radius_mean	0.170	Strong
texture_mean	-0.023	Medium
perimeter_mean	0.207	Strong
area_mean	0.177	Strong
compactness_mean	0.659	Strong
concavity_mean	0.522	Strong
points_mean	0.554	Strong
symmetry_mean	0.558	Strong
dimension_mean	0.585	Strong

compactness_mean

Statistics

Pearson Correlations

radius_mean	0.506	Strong
texture_mean	0.237	Strong
perimeter_mean	0.557	Strong
area_mean	0.499	Strong
smoothness_mean	0.659	Strong
concavity_mean	0.883	Strong
points_mean	0.831	Strong
symmetry_mean	0.603	Strong
dimension_mean	0.565	Strong



concavity_mean

Statistics

Pearson Correlations

radius_mean	0.677	Strong
texture_mean	0.302	Strong
perimeter_mean	0.716	Strong
area_mean	0.686	Strong
smoothness_mean	0.522	Strong
compactness_mean	0.883	Strong
points_mean	0.921	Strong
symmetry_mean	0.501	Strong
dimension_mean	0.337	Strong

points_mean

Statistics

Pearson Correlations

radius_mean	0.823	Strong
texture_mean	0.293	Strong
perimeter_mean	0.851	Strong
area_mean	0.823	Strong
smoothness_mean	0.554	Strong
compactness_mean	0.831	Strong
concavity_mean	0.921	Strong
symmetry_mean	0.464	Strong
dimension_mean	0.167	Strong

symmetry_mean

Statistics

Pearson Correlations

radius_mean	0.150	Strong
texture_mean	0.072	Strong
perimeter_mean	0.185	Strong
area_mean	0.153	Strong
smoothness_mean	0.558	Strong
compactness_mean	0.603	Strong
concavity_mean	0.501	Strong
points_mean	0.464	Strong
dimension_mean	0.479	Strong

dimension_mean

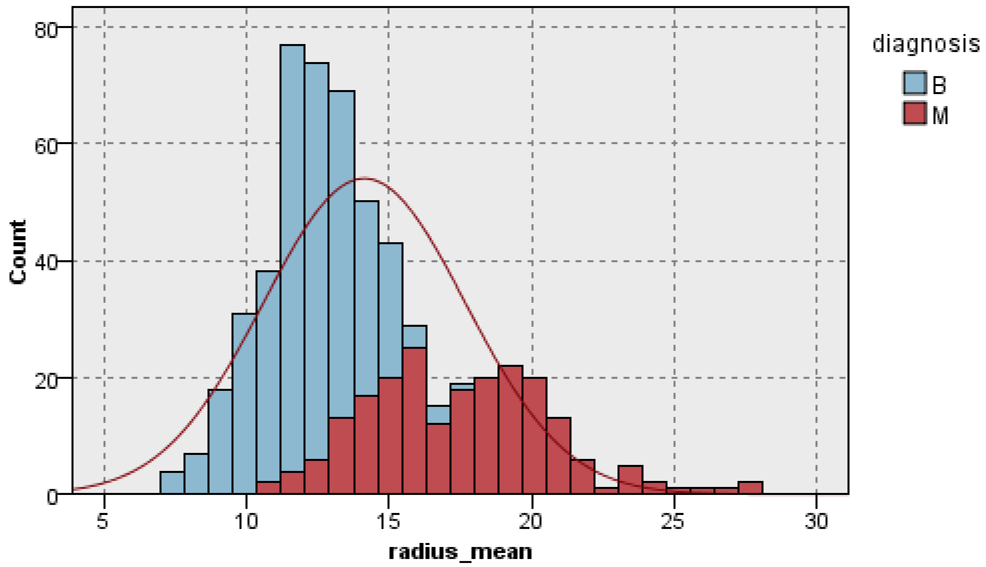
Statistics

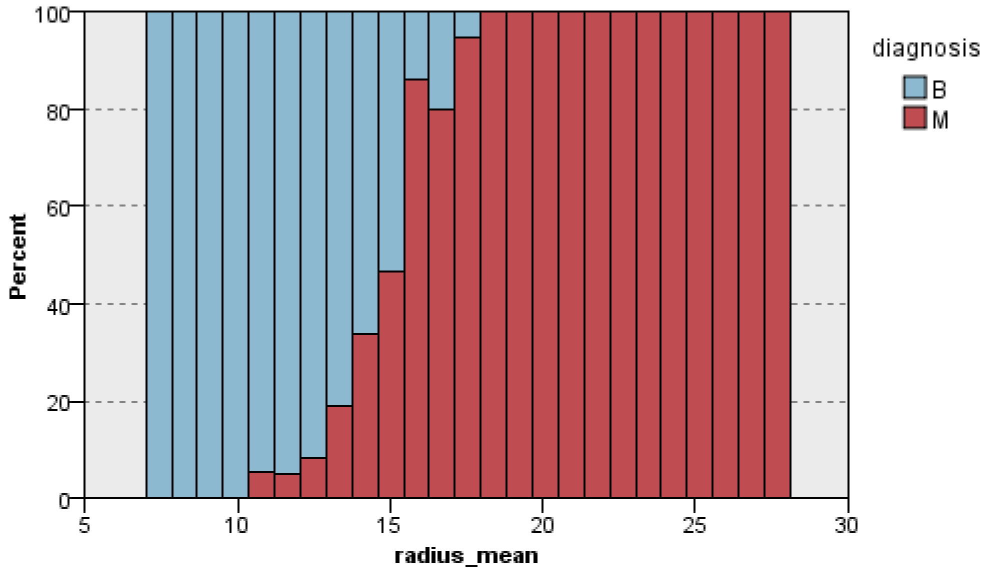
Pearson Correlations

radius_mean	-0.312	Strong
texture_mean	-0.076	Strong
perimeter_mean	-0.261	Strong
area_mean	-0.283	Strong
smoothness_mean	0.585	Strong
compactness_mean	0.565	Strong
concavity_mean	0.337	Strong
points_mean	0.167	Strong
symmetry_mean	0.479	Strong

From the above we can say that there is no corelated variables.

6. Construct a normalized histogram of radius_mean, with an overlay of the target variable diagnosis. Explain the results.





The above normalized histogram graph of the required attribute (radius_mean), the diagnosis is coded as "M" to indicate malignant, or "B" to indicate benign. The graph shows that radius_mean for (M) malignant are higher then as compared to (B) benign, these features relate to the shape and size of the cell nuclei.