

Statistics for Data Analytics

Continuous Assessment No.2

Q1.Perform analyses using correlation based/regression techniques (e.g. Multiple Regression, Logistic regression, Factor Analysis)

Ans.1 I have doing analyses on **multiple regression** as on this data linear regression will not work due to many predictors affecting on single dependent data.I have chosen the dataset on **World happiness report**.

Content

The scores are based this question, known as the Cantril ladder, asks the respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale.

The dataset includes country,region,happiness rank,happiness score,standard error,economy,family,health(life expectancy),freedom,trust,generosity,dystopia residual.

For more information follow this link:

<https://www.kaggle.com/unsdsn/world-happiness>

The dataset is publish by Sustainable Development Solutions Network released under CC0:Public Domain License

I have taken dataset of 158 rows from world happiness report which contains 12 columns regarding the people happiness.

I have used Happiness rank, family, health life expectancy, freedom, trust column in my regression analyses.

OBJECTIVE

My aim is to find the major reason of getting happiness. The column we can take for this happiness is **happiness rank** which will be the dependent variable (outcome) and it is calculated on the basis of family, health (life expectancy), freedom, trust (predictors-independent variables).

There are more than one predictor hence I cannot use linear regression hence I am using multiple regression.

How these factors will affect the happiness rank, this can be a very useful information to analyze and among which variable in a set of variables is the best factor of an outcome. It will help in finding the correlation between the factor and the outcome.

There are three number of multiple regression:

1. Standard or simultaneous
2. Hierarchical or sequential
3. Stepwise

I have used **standard regression** in which all the predictors are entered into the dataset simultaneously or I can say forced entry method where all variables are inserted into one go.

Preliminary test

There are some checks before multiple regression to meet the assumptions of it. Multiple regression is one of the fussier statistical techniques which makes a number of assumptions about the data and we cannot take the data if it gets violated.

1. Sample size

The issue is generalisability i.e small samples cannot generalise(repeated) with other samples. There is one equation that is $N > 50 + 8m$ (m means independent variables), I have 4 factors that means $N = 130$. So to meet the assumption I should have more than 130 cases and I have **157** cases which makes in favour of the condition (Pallant, J. 2013).

2. Multicollinearity and singularity

Represents the relationship between predictors. They both should not be there in multiple regression and if $r = .9$ or above then it shows multicollinearity and singularity occurs when one factor is combination of other independent variables.

My data doesn't contain these both quality hence I didn't violate this assumption (later on).

3. Outliers

Multiple regression is sensitive to outlier (very low or very high). There should not be any outlier in the case and in my data set there is no outlier hence I don't need to change or delete my outlier (Lind, D.A. and Marchal, W.G. et.al, 2011). It should be in between 3.3 to -3.3 if it goes beyond that limit then it's an outlier and I can show that I don't have any outlier from scatter plot later in the implementation.

4. Normality, linearity, homoscedasticity, independence of residuals

These refers to the distribution of scores and nature of the values. This can be checked from residuals scatterplot. The difference between obtained and the predicted dependent value scores (Lind, D.A. and Marchal, W.G. et.al, 2011)

. This checks:

1. Normality: residual should be normally distributed.

2. Linearity: it should have a straight-line relationship.

3. Homoscedasticity: variance of residual of predicted dependent scores should be same.

They all are checked and will be included in scatter plot and normal P-P plot of regression later on

Research Question-How well do the four measures (family, health, freedom, trust) predict happiness rank? How much variance in happiness scores can be explained by scores on these four measures?

I have used standard multiple regression. It means all of the independent variables being entered into the equation at once. The results will indicate how well the set of variables is able to predict happiness rank and it will also tell us how much different variance is each of the independent variables (family, health, freedom, trust) explains in the dependent variable.

Procedure:

1. Click on analyze>regression>linear.
 2. Click on continuous dependent variable(happiness rank),move to dependent box.
 3. Click on independent variable(family,health,freedom,trust) and move into independent box.
 4. method,enter should be there
 5. Click on statistics
 - 5.1 Tick box-estimates,Confidence intervals,model fit,descriptives,part and partial correlations and collinearity diagnostics.
 - 5.2 Residual section>tick casewise diagnostics & outliers outside.**(in my case there is no outlier hence will show all cases)**
 - 5.3 Continue
 6. Click options>include exclude cases pairwise.
 7. Click plots>
 - 7.1 Move ZRESID-Y BOX
 - 7.2 Move ZPRED-X BOX
 8. Click Save>distance-tick mahalanobis box (multivariate outlier) and cook's.
 9. Click ok.
- ((Kulas, J.T. 2008))

```
REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA COLLIN TOL ZPP
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT happiness_rank
/METHOD=ENTER Family Health_life_Expectancy Freedom Trust
/SCATTERPLOT=(*ZPRED ,*ZRESID)
/RESIDUALS NORMPROB(ZRESID)
/CASEWISE PLOT(ZRESID) OUTLIERS(3)
/SAVE MAHAL COOK.
```

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
happiness_rank	79.49367	45.754363	158
Family	.99587	.283297	158
Health_life_Expectancy	.63026	.247078	158
Freedom	.42861	.150693	158
Trust	.14342	.120034	158

Step 1:Check the assumptions

Multicollinearity:We can see the table of **correlations**,In my case (family,health,freedom,trust) correlate substantially with Happiness as they are more then **0.3** and there are more then two variables hence 0.7 is the exeptional case,we can take that(Lind, D.A. and Marchal, W.G. et.al ,2011).

Correlations

		happiness_rank	Family	Health_life_Expectancy	Freedom	Trust
Pearson Correlation	happiness_rank	1.000	-.731	-.736	-.557	-.372
	Family	-.731	1.000	.527	.448	.192
	Health_life_Expectancy	-.736	.527	1.000	.360	.248
	Freedom	-.557	.448	.360	1.000	.494
	Trust	-.372	.192	.248	.494	1.000
Sig. (1-tailed)	happiness_rank	.	.000	.000	.000	.000
	Family	.000	.	.000	.000	.008
	Health_life_Expectancy	.000	.000	.	.000	.001
	Freedom	.000	.000	.000	.	.000
	Trust	.000	.008	.001	.000	.
N	happiness_rank	158	158	158	158	158
	Family	158	158	158	158	158
	Health_life_Expectancy	158	158	158	158	158
	Freedom	158	158	158	158	158
	Trust	158	158	158	158	158

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	Trust, Family, Health_life_Expectancy, Freedom ^b	.	Enter

a. Dependent Variable: happiness_rank

b. All requested variables entered.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.865 ^a	.749	.742	23.221659

a. Predictors: (Constant), Trust, Family, Health_life_Expectancy, Freedom

b. Dependent Variable: happiness_rank

Collinearity diagnostics table pick up problems with multicollinearity that may not be evident, and it does violate this assumption.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	223.096	7.361		30.309	.000					
	Family	-65.682	8.166	-.407	-8.043	.000	-.731	-.545	-.326	.642	1.558
	Health_life_Expectancy	-80.680	9.004	-.436	-8.961	.000	-.736	-.587	-.363	.694	1.441
	Freedom	-50.560	15.600	-.167	-3.241	.001	-.557	-.253	-.131	.622	1.609
	Trust	-39.540	17.882	-.104	-2.211	.029	-.372	-.176	-.090	.745	1.341

a. Dependent Variable: happiness_rank

Collinearity Diagnostics^a

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	Family	Health_life_Expectancy	Freedom	Trust
1	1	4.528	1.000	.00	.00	.00	.00	.01
	2	.310	3.820	.01	.01	.02	.00	.79
	3	.078	7.624	.10	.00	.78	.18	.02
	4	.049	9.577	.48	.02	.04	.76	.16
	5	.034	11.485	.40	.96	.16	.05	.02

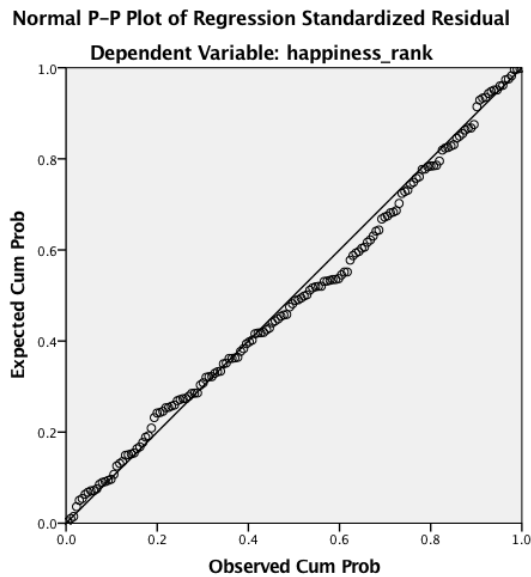
a. Dependent Variable: happiness_rank

We can see the tolerance value and VIF values **from coefficients table** and both are valid in this so we haven't violated this condition (not less than 0.10). Tolerance value is more than .10 and VIF is less than 10. Means there is no highly intercorrelated independent variables (Pallant, J. 2013).

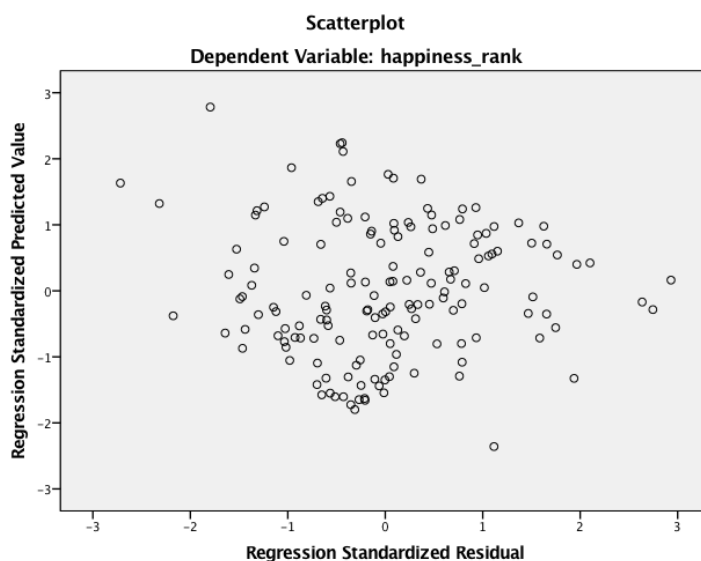
Outliers, Normality, Linearity, Homoscedasticity, Independence of Residuals

I can check this assumption from the **normal probability plot** that should make a straight line and my graph is almost making a straight line.

Charts



In the scatterplot, it is kind of making the roughly rectangularly distributed, with most of the values are in the center (around 0).



We don't have any outlier in the plot as all the values are **between 3.3 or less then -3.3**.

I have checked through mahalanobis distance and cook's distance also they are not exceeding the values from these too as maximum is checked from the residual table and it is below that hence we haven't violated this assumption (cooks distance below 1 and mahalanobis

distance is below 18.7).

Linear Regression: Save

Predicted Values

☒ Unstandardized

☐ Standardized

☐ Adjusted

☐ S.E. of mean predictions

Residuals

☐ Unstandardized

☐ Standardized

☐ Studentized

☐ Deleted

☐ Studentized deleted

Distances

☒ Mahalanobis

☒ Cook's

☐ Leverage values

Influence Statistics

☐ DfBeta(s)

☐ Standardized DfBeta(s)

☐ DfFit

☐ Standardized DfFit

☐ Covariance ratio

Prediction Intervals

☐ Mean ☐ Individual

Confidence Interval: 95 %

Coefficient statistics

☐ Create coefficient statistics

☒ Create a new dataset

Dataset name:

☐ Write a new data file

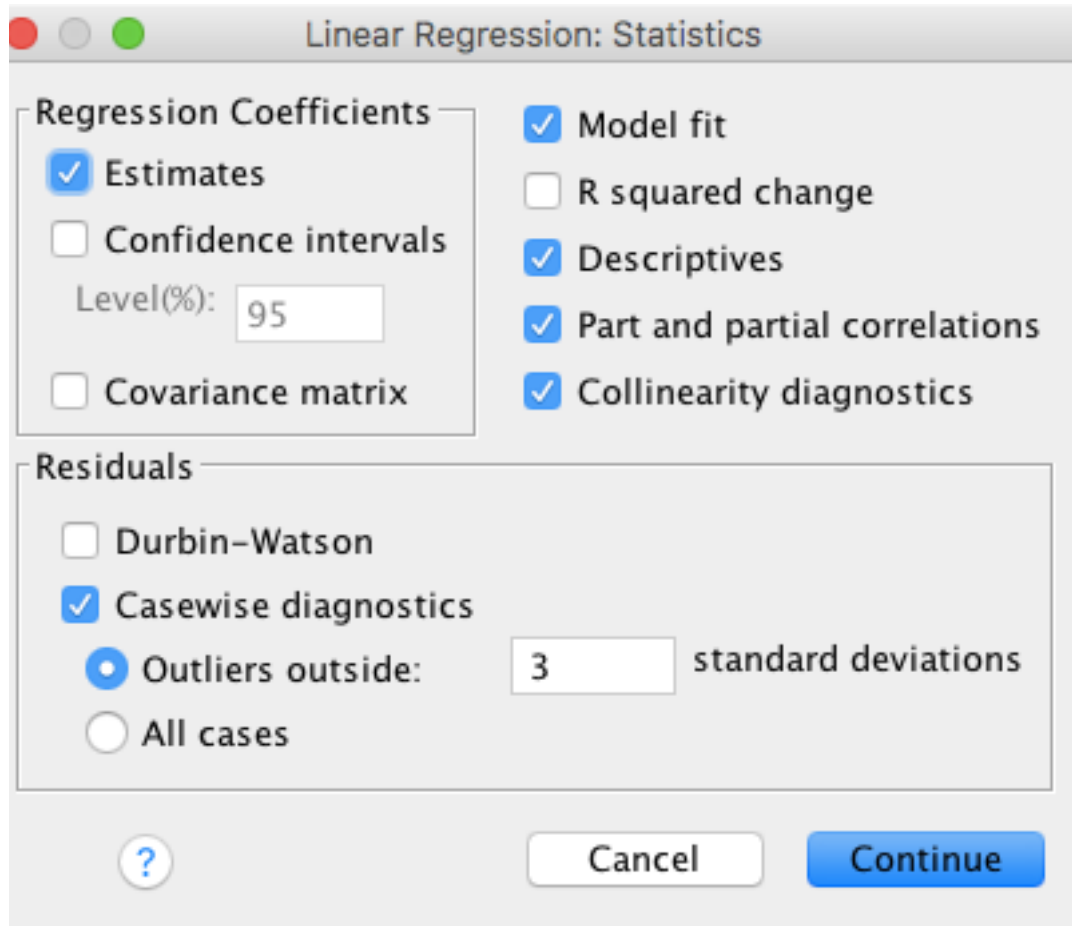
Export model information to XML file

☒ Include the covariance matrix

Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-13.90318	189.70020	79.49367	39.597412	158
Std. Predicted Value	-2.359	2.783	.000	1.000	158
Standard Error of Predicted Value	1.970	8.234	3.963	1.169	158
Adjusted Predicted Value	-16.75565	195.69774	79.48101	39.779457	158
Residual	-63.117168	68.059853	.000000	22.923933	158
Std. Residual	-2.718	2.931	.000	.987	158
Stud. Residual	-2.791	3.093	.000	1.006	158
Deleted Residual	-66.538765	75.783997	.012666	23.805221	158
Stud. Deleted Residual	-2.855	3.184	.001	1.014	158
Mahal. Distance	.136	18.748	3.975	3.118	158
Cook's Distance	.000	.217	.008	.022	158
Centered Leverage Value	.001	.119	.025	.020	158

a. Dependent Variable: happiness_rank



The image shows the 'Linear Regression: Statistics' dialog box in SPSS. It is divided into two main sections: 'Regression Coefficients' and 'Residuals'. In the 'Regression Coefficients' section, 'Estimates' is checked, 'Confidence intervals' is unchecked with a 'Level(%)' of 95, and 'Covariance matrix' is unchecked. In the 'Residuals' section, 'Durbin-Watson' is unchecked, 'Casewise diagnostics' is checked, and 'Outliers outside: 3 standard deviations' is selected with a radio button. At the bottom, there are buttons for '?', 'Cancel', and 'Continue'.

Section	Option	Status
Regression Coefficients	Estimates	Checked
	Confidence intervals	Unchecked
	Level(%)	95
	Covariance matrix	Unchecked
Residuals	Durbin-Watson	Unchecked
	Casewise diagnostics	Checked
	Outliers outside: 3 standard deviations	Selected
	All cases	Unselected
	Buttons	?, Cancel, Continue

In my case there is no outlier (above 3.0 or below -3.0) hence SPSS haven't made any **Casewise Diagnostics**.

Step:2 Evaluating the Model

The model summary tells how much variance in the dependent variable(happiness) is majoured by (family,health,freedom,trust).In this case my R square value is .749 that is equal to 74.9% of the variance in the happiness rank.This is a respectable result((Kulas, J.T. 2008).

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.865 ^a	.749	.742	23.221659

a. Predictors: (Constant), Trust, Family, Health_life_Expectancy, Freedom

b. Dependent Variable: happiness_rank

R should be less than 0.9 and from the snapshot it is clearly shown that $r=.8$.

To assess the significance, we can look at ANOVA table which tests the null hypothesis that multiple R is 0.

My significance is 0.000 which is less than 0.0005 (0.05).

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	246168.942	4	61542.235	114.127	.000 ^b
	Residual	82504.552	153	539.245		
	Total	328673.494	157			

a. Dependent Variable: happiness_rank

b. Predictors: (Constant), Trust, Family, Health_life_Expectancy, Freedom

Step 3:Evaluating each of the independent variables

The next thing which is important to know is which variables in the model contributed to the prediction of the dependent variable(happiness).

We can see this information in the outbox box like **coefficients**(Lind, D.A. and Marchal, W.G. et.al ,2011).See beta under standardised coefficients.Have to find which value in beta column is the largest among others ignoring negative sign,whichever has the highest beta coefficient makes the strongest unique contribution and it goes in descending order from unique to related(highest value to lowest).We have to see the

significance value to see the uniqueness. If the significance value of less than 0.5 then it makes a unique contribution (Pallant, J. 2013).

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1 (Constant)	223.096	7.361		30.309	.000					
Family	-65.682	8.166	-.407	-8.043	.000	-.731	-.545	-.326	.642	1.558
Health_life_Expectancy	-80.680	9.004	-.436	-8.961	.000	-.736	-.587	-.363	.694	1.441
Freedom	-50.560	15.600	-.167	-3.241	.001	-.557	-.253	-.131	.622	1.609
Trust	-39.540	17.882	-.104	-2.211	.029	-.372	-.176	-.090	.745	1.341

a. Dependent Variable: happiness_rank

Largest value in beta column is of health hence it means that it is the strongest unique contribution to dependent variable and trust is the lowest.

In this case every predictor has less than 0.5 significance value hence making them significant unique contribution of dependent variable.

In the part column, all value should be multiplied by itself then it will show the variance in the happiness scores.

Family=10.5%

Health=13.17%

Freedom=1.7%

Trust=0.8%

These are the percentage of the variance in the happiness scores.

The result of the analyses according to the question which includes predictors like family, health, freedom and trust makes **74.9%** of the variance in the happiness scores.

Health makes **the largest unique contribution** among others and trust is **the least unique contribution** among others which gives happiness.

Hence a person is getting happy mainly by health (life expectancy) followed by family, freedom, trust.

Q2.Perfrom analysis and it should use a technique to compare observations between groups(eg.ANOVA,MANOVA,suitable non-parametric tests)

Ans.2 For the second question of analysis, I will be performing **one-way ANOVA** between groups in the data set.

Abstract

I have **taken video games sales data** of 149 games of different genre and ranks which contains 11 colums regarding their sales data in different regions of the world.I am using anova instead of t test because there are more then 2 number of groups in independent variable.

This data is published by Gregory Smith's by web scarping of VGCHArtz video games sales from Metacritic which is updated again before 3 months.

The data set contains a Rank, Name, Platform, Year_of_Release, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales.

To get more information follow the link:

<https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings>

OBJECTIVE

The objective behind this analysis is to find which publisher of the game like nitendo,advision have the maximum sales in **North America**.

(I am calculating the data on **North America sales**)

Labelling of analyses are as follows:

Publisher:name of the company which published the game

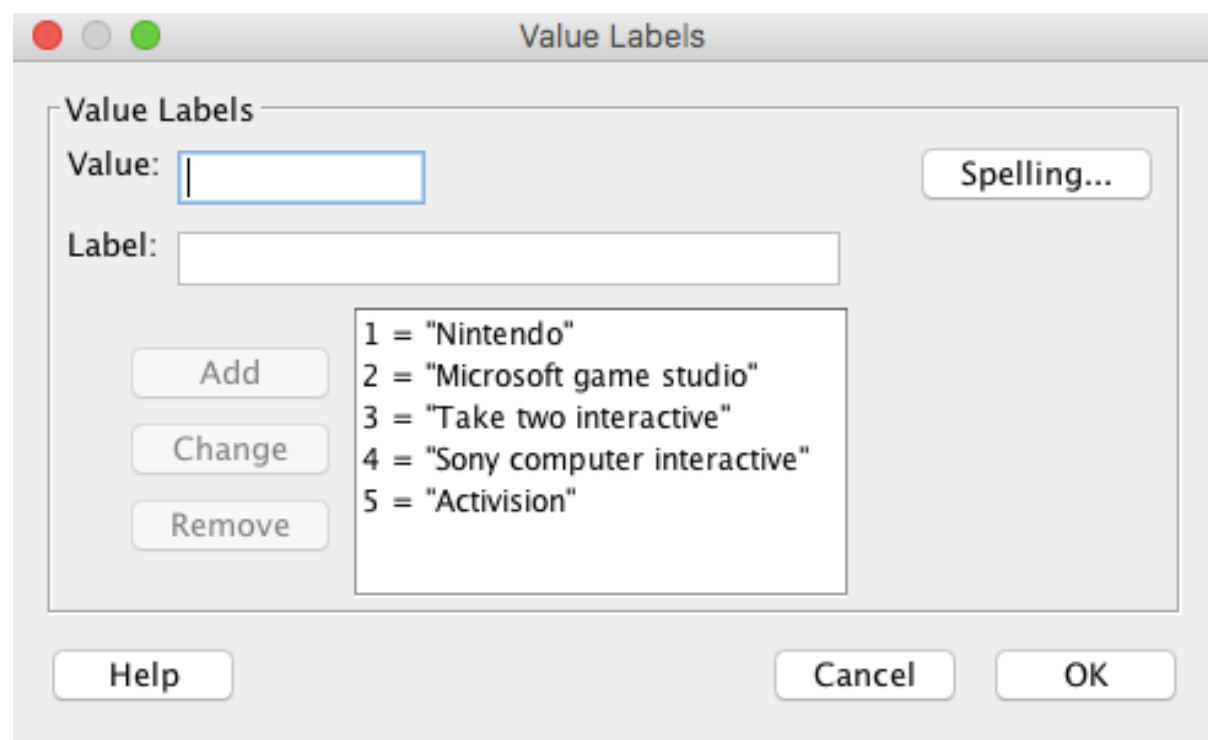
1-"Nintendo"

2-"Microsoft game studio"

3-"Take two interactive"

4-"Sony computer interatctive"

5-"Activision"



From this group,I can find which has the maximum sales in North America which can be find by histogram.

But to find the variance between the groups and to know the information into detail,I have to drill-down by using one-way ANOVA.

North America sales is **dependent variable** as due to the publication of these companies.

Publisher name is **independent variable** as it is not dependent on anyone to change the strata of the data,so it is a fixed factor.

Preliminary tests

There are certain assumptions before analysing the anova to get perfect results as if it gets violated we cannot perform anova on that set and we have to switch to non-parametric tests.

Anova is just like T test but we cannot perform more complications in T test (2 groups maximum) and anova(2 or more groups) is best suited to find variance between the group.

Assumptions:

There are 6 assumptions that should be met :

1.Before performing analysis,my dependent variable should be interval level or ratio level with a continuous scale.

2. Two independent variables should have two or more categorical, independent groups- I have 5 independent groups with one dependent variable.

3.My **independent samples** are assigned to treatment combinations.There are no repeated subjects in any factor integration.

4.There should be no significant outlier.

5. Dependent data should be approximately normally distributed for each category of the fixed factor (**robustness** need to be checked).

6. There should be homogeneity of variance (Lind, D.A. et al, 2011)

To check these assumptions:

1. Calculate via SPSS **Levene's test** to test the homogeneity of variance and if we get significance **value p more than 0.05** then we haven't violated the assumption of ANOVA.

2. Calculate **ANOVA significance** from ANOVA table and check the value of significance p.----- (F) in table

If it's **less than 0.05** then we haven't violated the assumption.

3. Calculate the value of **Welch and Brown-Forsythe** and check its significance, if we get **less than 0.05** then we haven't violated the **assumption**.

(Pallant, J. 2013)

I have used **histogram** to illustrate the objectives and points more clearly.

For this I have done:

1. Go to SPSS software

2. Import the file and change the types to numeric, and change health to scale measure and add values

3. Go to graphs

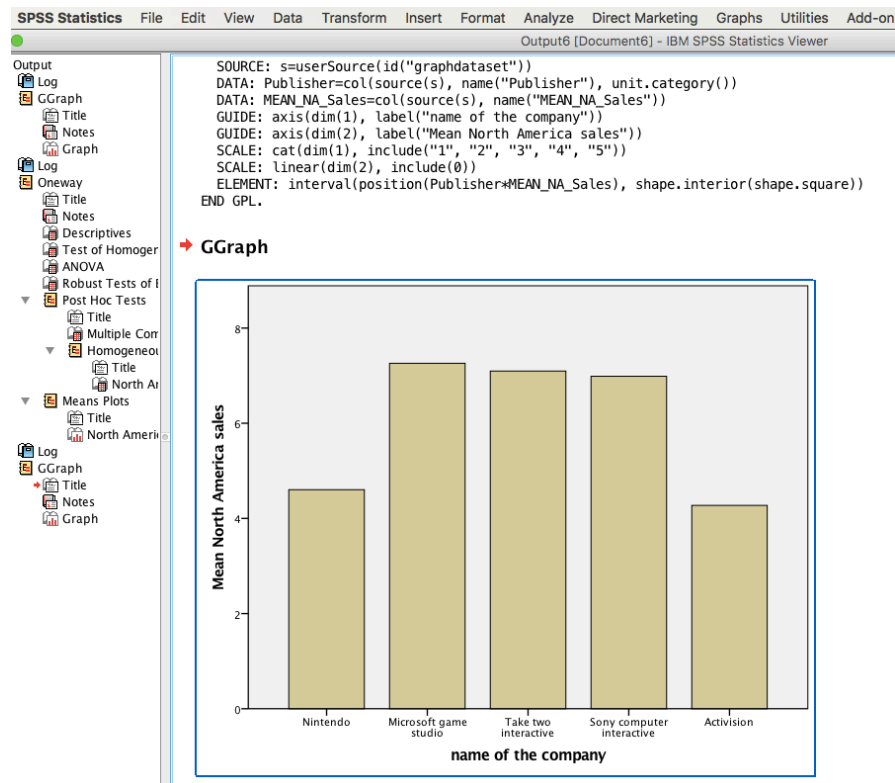
4. Click chart builder

5. Choose bar from the bottom left column and drag into the space.

6. Allocate the variables into the field

7. Click element properties then click bar 1 and mean then click apply and ok

8. Graph will be shown in the output window.



DATA IS NOT BELL-SHAPED PROPERLY BUT THE DATA WE ARE TAKING IS MORE THEN 30.SO WE CAN TAKE ITS NORMALLY DISTRIBUTED.

Here,**publisher** is on X axis,**North America sales** is on Y axis.

It is calculating the video games sales from five different types of publisher.It is showing microsoft game studio does the **maximum** sales in North America and Activision is doing **least**(millions).

I will be performing one-way anova on these variables to check the variance between and within the terms including the effect.

Steps:

- 1.Go to analyze on the menu bar
- 2.Click compare means
- 3.Select one-way anova
- 4.Import dependent and independent variables
- 5.Select post-hoc-tukey test

6. Select options-mark descriptive, homogeneity of variance, brown-Forsythe, welchm, means plot

7. Click ok

Now we should check the homogeneity of variances by using Levene's test.

```
ONEWAY NA_Sales BY Publisher
  /STATISTICS DESCRIPTIVES HOMOGENEITY BROWNFORSYTHE WELCH
  /PLOT MEANS
  /MISSING ANALYSIS
  /POSTHOC=TUKEY BONFERRONI ALPHA(0.05).
```

→ Oneway

Descriptives

North America sales								
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Nintendo	62	4.60	5.119	.650	3.30	5.90	0	41
Microsoft game studio	20	7.26	5.553	1.242	4.66	9.86	1	27
Take two interactive	28	7.10	4.971	.939	5.17	9.03	3	23
Sony computer interactive	22	6.99	6.121	1.305	4.27	9.70	1	29
Activision	17	4.27	2.479	.601	3.00	5.55	1	9
Total	149	5.74	5.189	.425	4.90	6.58	0	41

Test of Homogeneity of Variances

North America sales			
Levene Statistic	df1	df2	Sig.
1.515	4	144	.201

Levene statistic is used to check if the variance is the same for different groups.

If the $p > 0.05$ the assumption of variance is not violated as I mentioned above. The significant level is 0.201 which is much more than 0.05 then it is not violated.

In case if it gets violated we can perform **Robust Tests of Equality of Means** (Kulas, J.T. 2008).

There is no sign of interaction between variables (independency).

Random sampling is done to complete a random sample of 27 entries which is 19 percent of total sample of 149 entries.

OUTPUT:

```
FREQUENCIES VARIABLES=NA_Sales Publisher
/STATISTICS=STDDEV SEMEAN MEAN
/ORDER=ANALYSIS.
```

➔ Frequencies

Statistics

		North America sales	name of the company
N	Valid	27	27
	Missing	0	0
Mean		4.45	1.33
Std. Error of Mean		1.439	.141
Std. Deviation		7.479	.734

name of the company

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nintendo	22	81.5	81.5	81.5
	Microsoft game studio	1	3.7	3.7	85.2
	Take two interactive	4	14.8	14.8	100.0
	Total	27	100.0	100.0	

Now look at the **Output** of analyses including Levene test:

```

ONEWAY NA_Sales BY Publisher
  /STATISTICS DESCRIPTIVES HOMOGENEITY BROWNFORSYTHE WELCH
  /PLOT MEANS
  /MISSING ANALYSIS
  /POSTHOC=TUKEY BONFERRONI ALPHA(0.05).

```

➔ Oneway

Descriptives

North America sales

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Nintendo	62	4.60	5.119	.650	3.30	5.90	0	41
Microsoft game studio	20	7.26	5.553	1.242	4.66	9.86	1	27
Take two interactive	28	7.10	4.971	.939	5.17	9.03	3	23
Sony computer interactive	22	6.99	6.121	1.305	4.27	9.70	1	29
Activision	17	4.27	2.479	.601	3.00	5.55	1	9
Total	149	5.74	5.189	.425	4.90	6.58	0	41

Test of Homogeneity of Variances

North America sales

Levene Statistic	df1	df2	Sig.
1.515	4	144	.201

Descriptive table:

(SALES IS IN MILLION)

Numbers are correctly into the table-149

Look at the mean of nintendo publisher it sells 4.60 million video games in North America, Microsoft game studio sells 7.26 million, take two interactive sells 7.10, Sony sells 6.99 million, activision sells 4.27 million and we have found the same information from the histogram that **microsoft** sells the maximum and **activision** is least, standard deviation is different in each group at least 7.1 is different then 2.5.

This indicates that at least one comparison of variance is significant which follows the assumptions.

ANOVA analysis:

ANOVA

North America sales

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	265.285	4	66.321	2.567	.041
Within Groups	3720.218	144	25.835		
Total	3985.503	148			

This table gives both between-groups and within-groups sums of squares, degrees of freedom etc.

The main thing in the table is significance ,here the P value is 0.041 which is lesser than 0.05 that indicates there is a significant difference among the mean scores on the dependent variable North America Sales for various Publisher groups. The means for each group are given in the **Descriptives** table.

Further we can look at the results of the post-hoc tests later on.

We need to check the robust tests of Equality of means to check the assumptions that we met or not.**Output:**

Robust Tests of Equality of Means

North America sales

	Statistic ^a	df1	df2	Sig.
Welch	2.869	4	55.687	.031
Brown-Forsythe	2.503	4	92.696	.048

a. Asymptotically F distributed.

Robust case have two test:

1.Welch

2.Brown-Forsythe

Robust test accommodates heterogeneous variances. Welch statistic is significant as its p value (0.31) is less than 0.05. The degrees of freedom in numerator is 4 and in denominator is 4.341. Brown-Forsythe statistics is also significant as its p value (.048) is less than 0.05. These two tests are measured to check the assumption and this indicates that we will reject the null hypothesis.

Result from Robust test = Significant difference in means.

Multiple Comparisons:

This table should be seen only if anyone found a significant difference in overall ANOVA. The table below, **Multiple Comparisons**, shows which groups differed from each other. The Tukey-post hoc test is the preferred one so I am using this test. There are no asterisks (*) next to the values listed; this means that there is a significant difference in the groups. This is an extension of ANOVA table (Marini, F. and de Beer, D. et al, 2017).

Post Hoc Tests

Multiple Comparisons

Dependent Variable: North America sales

Tukey HSD

(I) name of the company	(J) name of the company	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Nintendo	Microsoft game studio	-2.657	1.310	.258	-6.28	.96
	Take two interactive	-2.495	1.160	.204	-5.70	.71
	Sony computer interactive	-2.386	1.264	.329	-5.88	1.11
	Activision	.329	1.395	.999	-3.52	4.18
Microsoft game studio	Nintendo	2.657	1.310	.258	-.96	6.28
	Take two interactive	.162	1.491	1.000	-3.96	4.28
	Sony computer interactive	.271	1.574	1.000	-4.08	4.62
	Activision	2.986	1.680	.391	-1.66	7.63
Take two interactive	Nintendo	2.495	1.160	.204	-.71	5.70
	Microsoft game studio	-.162	1.491	1.000	-4.28	3.96
	Sony computer interactive	.110	1.451	1.000	-3.90	4.12
	Activision	2.824	1.566	.376	-1.50	7.15
Sony computer interactive	Nintendo	2.386	1.264	.329	-1.11	5.88
	Microsoft game studio	-.271	1.574	1.000	-4.62	4.08
	Take two interactive	-.110	1.451	1.000	-4.12	3.90
	Activision	2.715	1.645	.468	-1.83	7.26
Activision	Nintendo	-.329	1.395	.999	-4.18	3.52
	Microsoft game studio	-2.986	1.680	.391	-7.63	1.66
	Take two interactive	-2.824	1.566	.376	-7.15	1.50
	Sony computer interactive	-2.715	1.645	.468	-7.26	1.83

Homogeneous Subsets

North America sales

Tukey HSD^{a,b}

name of the company	N	Subset for alpha = 0.05
		1
Activision	17	4.27
Nintendo	62	4.60
Sony computer interactive	22	6.99
Take two interactive	28	7.10
Microsoft game studio	20	7.26
Sig.		.252

Means for groups in homogeneous subsets are displayed.

- Uses Harmonic Mean Sample Size = 24.258.
- The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

We have significance level which is **greater than 0.05** in each group, it doesn't reject the null hypothesis and tell which groups are not significantly different.

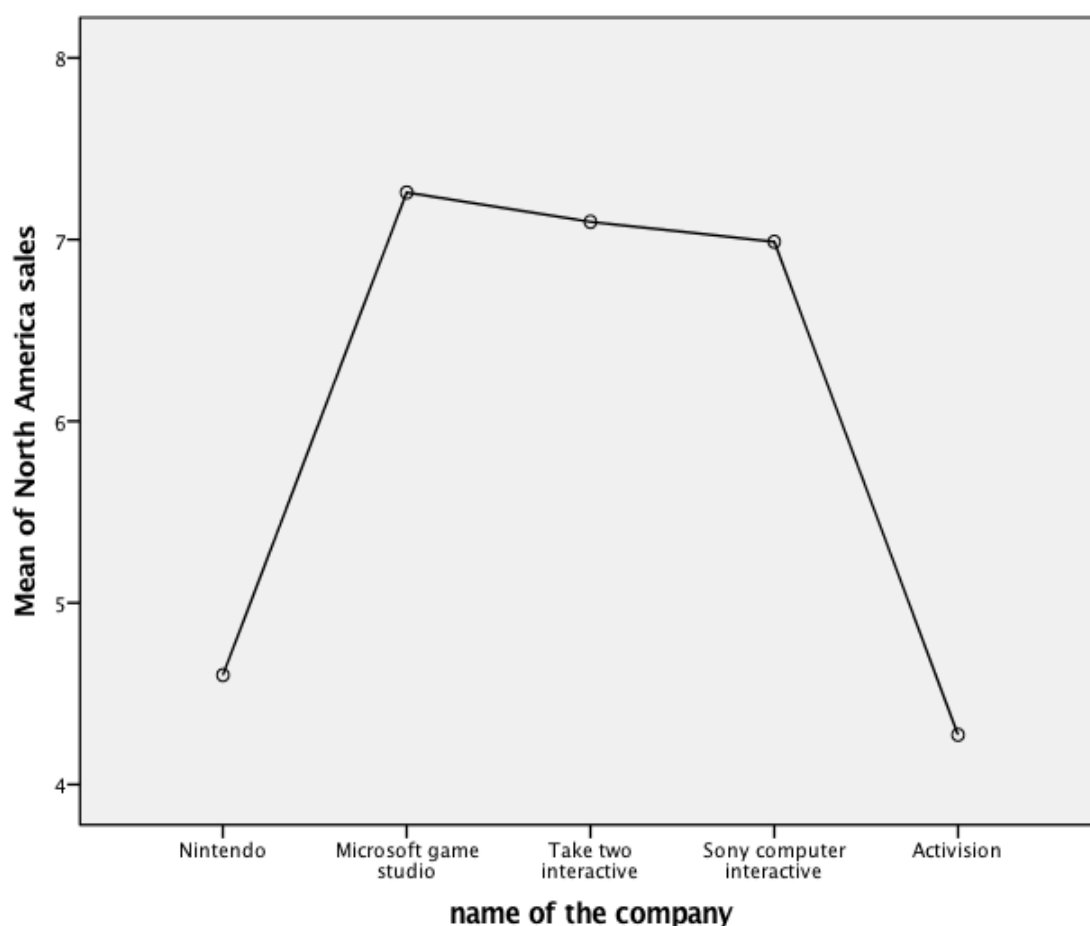
Means Plot

The plot gives an easy way to compare the mean scores for the various groups.

X axis-Publisher(name of the company who published the game)

Y axis-North America sales

Means Plots



There is a significant difference somewhere among the mean scores on the dependent variable **North America Sales** for different

Publisher groups.

Microsoft game has the maximum that is between 7-8 million sales and activision has the minimum sales between 4-5.

Result for analyses:

The analysis was conducted to find or know the comparisons between groups and to know which Publisher Group have the maximum sales in North America. I chose the one-way ANOVA technique and found out the Microsoft game has the maximum sales and the activision has the minimum(Cortinhas, C. & Black, K., 2012).

The results of the one-way between-groups analysis of variance with post-hoc tests could be presented as follows:

A one-way between-groups analysis of variance was conducted to explore the video game publisher group on North America sales. It was divided into five groups (Nintendo, Microsoft game studio, take two interactive, sony computer interactive, activision). There was a statistically significant difference at the $p < .05$ level in scores for the five groups [$F(4, 148) = 2.567, p = .041$]. It doesn't reject the null hypothesis and tells groups aren't significantly different (Pallant, J. 2013, p.220). Post-hoc multiple comparisons using the Tukey test indicated that the mean score for nintendo ($M = 4.60, SD = 5.119$), microsoft game ($M = 7.26, SD = 5.6$), take two interactive ($M = 7.10, SD = 5$), sony computer ($M = 7, SD = 6.121$) and activision ($M = 4.27, SD = 2.5$) all are significantly different.

References:

- Cortinhas, C. & Black, K. 2012, *Statistics for business and economics*, 1st European edn, Wiley, Chichester.
- Kulas, J.T. 2008, *SPSS essentials: managing and analyzing social sciences data*, Jossey-Bass, Chichester; San Francisco, Calif;.
- Lind, D.A., Marchal, W.G. & Wathen, S.A. 2011, *Basic statistics for business & economics*, 7th, International student edn, McGraw-Hill Irwin, New York.
- Marini, F., de Beer, D., Walters, N.A., de Villiers, A., Joubert, E. & Walczak, B. 2017, "Multivariate analysis of variance of designed chromatographic data. A case study involving fermentation of rooibos tea", *Journal of Chromatography A*, vol. 1489, pp. 115-125.
- Pallant, J. 2013, *SPSS survival manual: a step by step guide to data analysis using IBM SPSS*, 5th edn, McGraw-Hill, Maidenhead.