# Testing Gladwell Malcolm theory about hockey player birthday.

Using descriptive statistics and visualization

Data File: HockeyPlayers.sav

Pre-processing:

Birth\_Country\_Trimmed, Position\_Trimmed: fields stripped from spaces and quotation marks

## a)Graph to show whether Mr. Gladwell’s theory is true.

1. Filter data for Canada only : Data->Select->if Birth\_Country\_Trimmed = ‘Canada’
2. Create bar graph Graph->Chart Builder: Select simple bar graph Set x – Birth\_month

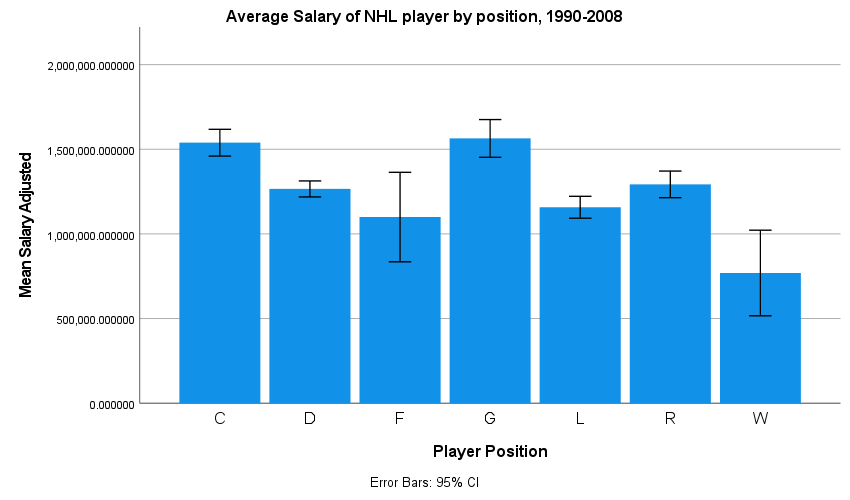
Chart, bar chart, histogram

Description automatically generated

1. *This graph displays that number of players in NHL born earlier in a year (Jan-Jun) is higher then number of players born at the end of the year. Nov and December demonstrate the lowest number of players, almost half less.*

## b. Highest paid position

1. Filter out records with missing data: Data->Select-> Position\_Trimmed<> "."
2. Create bar graph: Graphs-> Chart builder-> Simple bar graph-> x-axes – Position\_trimmed, y-axes- mean of Salary, checkbox Display error bars.



Graph shows that on average the highest paid position is G, goalie.

## Are Canadian born players, on average, paid higher than players born in other countries?

Group: Canadians, other -> calculable variable:

1. Create Boolean variable “Canadian” : Transform->Record into different -> Birth\_County\_Trimmed into Canadian, old ‘Canada’ ->1, Else->0
2. Filter out the records with empty country of birth: transfor->select->Birth\_Country\_trimmed<>’.’
3. Analyse->Mean-> set up

Graphical user interface, application

Description automatically generated

4. result: descriptive average

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
| Salary mean report | | | |
| Canadian | Mean | N | Std. Deviation |
| 0 Not Canadian | 1499147.14959521 | 4002 | 1717739.800378695 |
| 1 Canadian | 1226435.40684573 | 5957 | 1493423.775600573 |
| Total | 1336023.95935938 | 9959 | 1592916.375542236 |

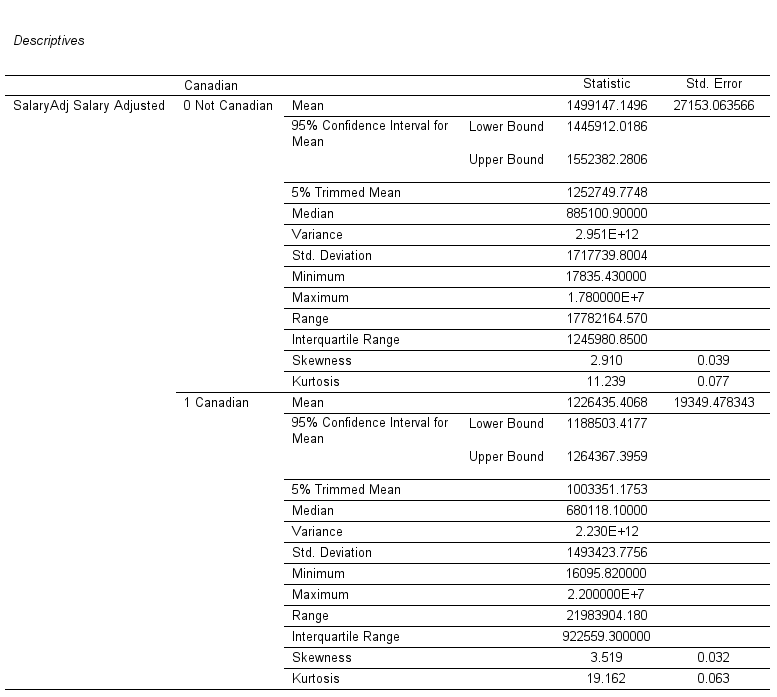
Descriptive statistics indicates that Canadian-born player are paid lower then non-Canadian-born

There are outliers that influence the means, so I will check full descriptive statistics

1. Remove splitting – analyse all cases
2. Create descriptive summary

Graphical user interface, application

Description automatically generated



Conclusion: according to surveys collected 1990-2008, on average salary of Canadian-born players lower then non-Canadian born players.

I would also like to compare means though the years (data for survey 2004/05 are missing)

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

Because salaries are skewed, I created both average(Mean) and median plots to see the difference in average for Canadian-born and non-Canadian players. Looks like Canadian-born are paid less on average, especially in the last two surveyed seasons even though one Canadian-born is the highest-paid player.

## d)Analyze the distribution of salaries.

I analyze salaries for one year (2007/2008) time=17, because for the whole 18 years, they will be skewed because of the inflation and growth of the salaries while years past.

1. data-> select cases-> if time=17
2. Normality test: analyze->descriptive statistics->explore
3. Set Dependent list -> salary
4. Statistics: check outliers
5. Plots: histogram
6. Options: exclude cases pairwise

Result:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Descriptives for salary adjusted for 2008/09 season* | | | | |
|  | | | Statistic | Std. Error |
| SalaryAdj Salary Adjusted | Mean | | 1831929.48952096 | 67142.524611566 |
| 95% Confidence Interval for Mean | Lower Bound | 1700093.33188299 |  |
| Upper Bound | 1963765.64715893 |  |
| 5% Trimmed Mean | | 1623365.32268796 |  |
| Median | | 1000000.00000000 |  |
| Variance | | 3011423232291.486 |  |
| Std. Deviation | | 1735345.277543200 |  |
| Minimum | | 475000.000000 |  |
| Maximum | | 1.000000E+7 |  |
| Range | | 9525000.000000 |  |
| Interquartile Range | | 1850000.000000 |  |
| Skewness | | 1.798 | 0.095 |
| Kurtosis | | 3.194 | 0.189 |

Standard error =0.05,

Analysis for normality using skewness and Kurtosis will be done according to documentation (IBM Documentation ,2022).

Skewness z-value=1.798/0.095=18.92 >>2 high right (positive) skewness

There are much more players with relatively small salaries (left side of the distribution) than players with big salaries. Mean is going to be compromised by small number of players who are not representative to the population, but their salaries are big enough to influence the mean (as an average).

Kurtosis z-value = 3.194/0.189 = 16.89 >>2 leptokurtic

The distribution has a high peak with more then normal number of cases that are far from mean(outliers)

Normality test:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Tests of Normality* | | | | | | |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
| Statistic | df | Sig. | Statistic | df | Sig. |
| SalaryAdj Salary Adjusted | 0.217 | 668 | 0.000 | 0.762 | 668 | 0.000 |
| a. Lilliefors Significance Correction | | | | | | |

Both tests indicate

P=0.00 => reject normality

Chart

Description automatically generated

Chart, line chart

Description automatically generated

Data are not following trend line in Normal QQ plot

Conclusion: reject normality of salaries.

As expected, a distribution of salaries is skewed right (positively skewed). The right tail (larger values) is much longer than the left tail (small values).There are just a few observations that are much larger than the rest.

# Speeding data analysis..

**File SpeedingData.xlsx**

1. Using one formula (no filters or pivot tables), find the average speed on the days when the roads are dry and there is no snow and the visibility is good and there is not congestion.

Formula: {=AVERAGE(IF('Sample 2'!C8:C10007+'Sample 2'!D8:D10007+'Sample 2'!E8:E10007+'Sample 2'!F8:F10007=0,'Sample 2'!B8:B10007,""))}

Graphical user interface, application, table, Excel

Description automatically generated

1. Set up appropriate bins for grouping the speed data

10,000 data points

Bin width = 10

1. Using an array function, find the frequency for speeds in each of the bins you’ve created, and create a histogram of the results.

See excel file

I defined name speed for speed array of data

Interval E3:E19 contains bins

={FREQUENCY(speed,E3:E19)}

I used Histogram tool of DataAnalysis add-on to create histogram.

Graphical user interface, chart, histogram

Description automatically generated

# Students grades and self-efficacy survey.

You have been hired to investigate the relationship between a student’s grade goal and self-efficacy (both ‘specific’ and ‘general’) at the outset of an MBA programme and their perceptions of social support and their actual average grade at the end of the programme (approximately 18 months later). Age and gender of the students were obtained as control variables. Details of the variables and their measurement are provided in Appendix 2.

**Tasks:**

1. Create a variable **ssupport** that is a summated scale, comprising seven items (see Appendix 2). Undertake a reliability analysis for this scale and comment on the acceptability of the Cronbach’s alpha coefficient.

|  |  |  |  |
| --- | --- | --- | --- |
| 1: Reverse negatively worded items | | | |
| Variable | Position | Label |  |
| Pplquit | 8 | People on the MBA often think about quitting | N |
| Quitting | 9 | I frequently think of quitting the MBA | N |
| Enjoy | 10 | I enjoy my studies on the MBA Programme | P |
| Relsfac | 11 | I feel satisfied with my relations with faculty and tutors | P |
| Support | 12 | I get sufficient support from the university | P |
| Stressful | 13 | The MBA is very stressful | N |
| Interfere | 14 | The stress involved with the MBA doesn't interfere with my academic progress | P |

Create new variable pplquit\_r,quitting\_r, stressful\_r

Transform->Recode Into Different Variables

2.Create ssupport: Transform->Compute

ssupport=pplquit\_r+quitting\_r+enjoy+relsfac+support+stressfu\_r+interfer.

Changed variable ssupport type to scale, set decimals to 0

1. Reliability analysis for this scale

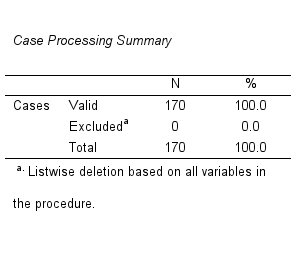
Analyze->Scale->Relibility Test

Graphical user interface, application

Description automatically generatedGraphical user interface, text, application, email

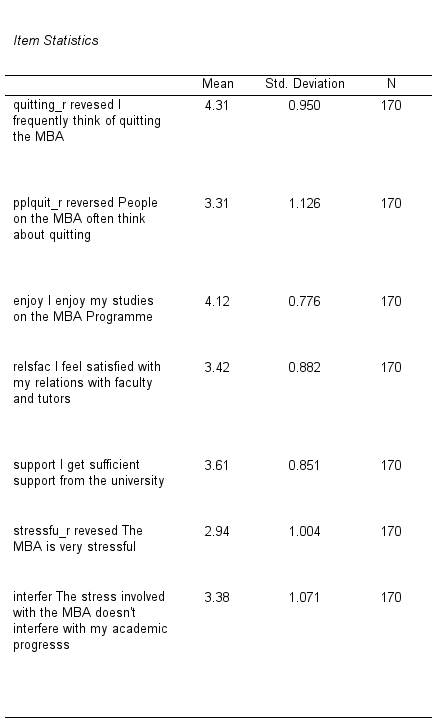
Description automatically generated

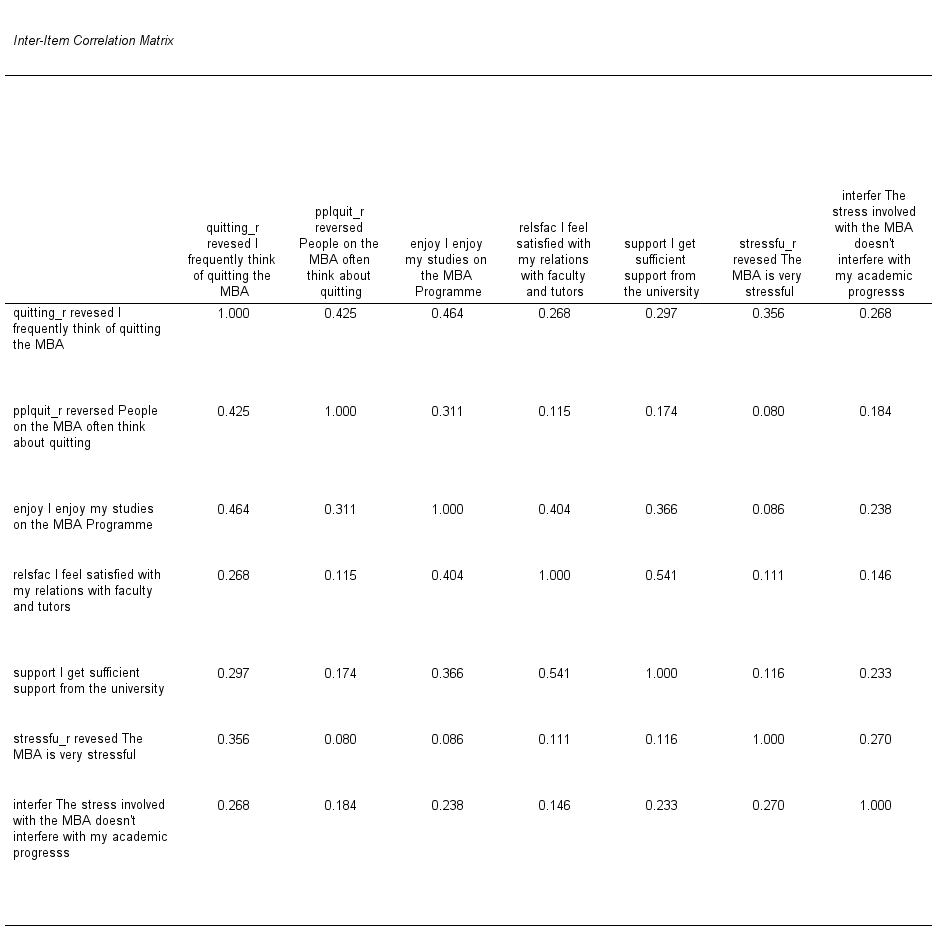
Result:

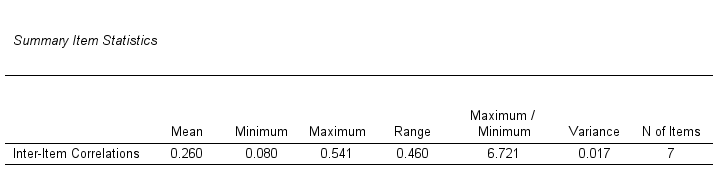


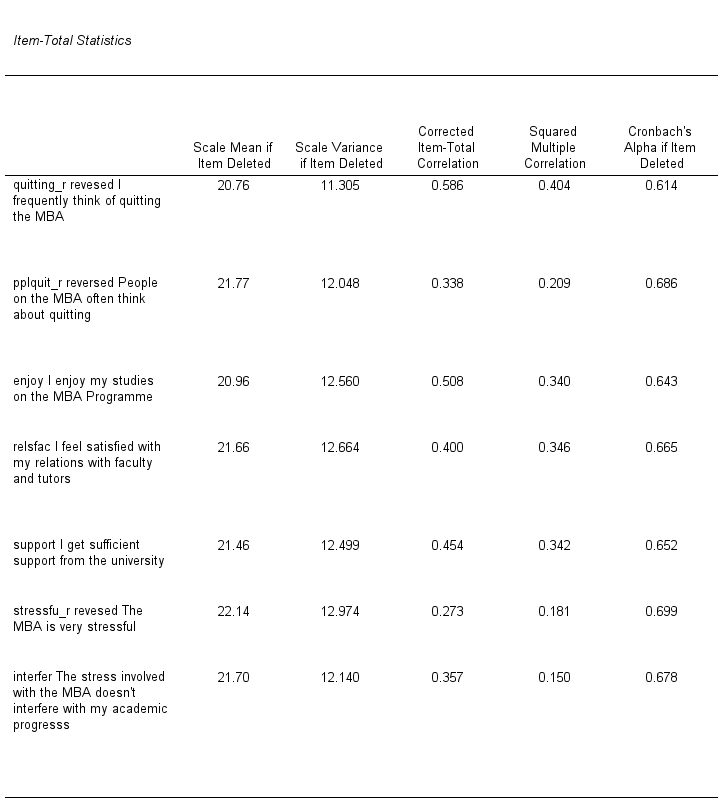
Table

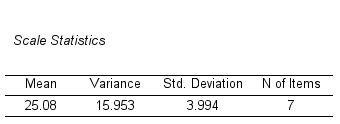
Description automatically generated











1. comment on the acceptability of the Cronbach’s alpha coefficient.

Cronbach’s Alpha value is equal 0.697. Because it is fewer then 0.7, the score is not satisfying. Reversed stressfu has corrected score 0.273<0.3 and if we remove it from the score, it will slightly increase Cronach’s Alpha value to 0.699.

Another comment, we have 7 items in the scale, which is fewer then 10. According to our textbook (Pallant,2020), if there are less then 10 items in the score, it is sometimes difficult to get decent value. We can use mean inter-item correlation of 0.260 (0.150<0.260<0.5) to support that our score is consistent.

References:

1. IBM Documentation (2022). [Summarize Statistics - IBM Documentation](file:///C:\Humber\BIA\Data%20Analytics%20Tools\Assignment1\Summarize%20Statistics%20-%20IBM%20Documentation) retrieved from <https://www.ibm.com/docs/en/spss-statistics/SaaS?topic=summarize-statistics>
2. Pallant, Julie (2020). SPSS Survival Manual A Step by Step Guide to Data Analysis Using IBM SPSS, 7th edition.