Terminal Assignment Based Assessment

Ross Currid
x20147147@student.ncirl.ie

***Abstract*—This assignment contains two analyses; a principal component analysis (PCA) of Car Types in America and a Two-Way ANOVA (with interaction/replication) analysis of a manually created dataset relating to a pharmaceutical company. You will also find brief literature reviews pertaining to the use of Factor Analysis and ANOVA.**

Keywords—Eigenvalue, var, PCA, ANOVA

I. PCA

A. Introduction

This dataset relates to Cars in America and their attributes as well as Cost & Sale Measures. As PCA works best with numerical variables, certain measures have been removed from my dataset as illustrated in part A of section III. I have elected to investigate the role of Car Types (Minivan, Pickup, Regular, Sports, SUV, Wagon) in the dataset and determine how much of the data can be assigned or related these variables.

The purpose for doing this is to reduce the dimensionality of the dataset and identify how many dimensions are of little relevance to the data so as to prioritise how to go about further analysis of the data. In this sense, I am hoping to derive a small number of components that can account for the variability found in the dataset measures.

B. Pre-Processing

I changed each variable in my target column (Car Type) to 1, 2, 3, 4, 5 and 6 and subsequently converted the column type to numeric.

The following code was used to loop over my dataset and replace column NA's with their means:

"NA2mean <- function(x) replace(x, is.na(x), mean(x, na.rm = TRUE))
replace(df, TRUE, lapply(df, NA2mean))"

Two other category columns were removed prior to this.

C. Background

Factor Analysis has long been used as a method to reduce a large number of variables into fewer numbers of factors and its use has become more widespread with the growth of data science and computing capabilities (Williams et. al, 2010).

Factor Analysis has its root in psychometrics with Charles Spearman, a psychologist being the first to discuss common factor analysis in 1904. Since then, it has taken on many new shapes and forms, most notably helped in this regard by Louis Thurstone in his seminal book, *The Vector of the Mind* (1935).

It can be broken into two categories, exploratory and confirmatory factor analysis. In this sense, exploratory factor analysis could be considered qualitative and confirmatory factor analysis quantitative. An example of the former would be use of the K-Means model while an example of the latter is Structural Equation Modelling. As Jamie DeCoster succinctly outlines in his paper, an 'Overview of Factor Analysis' (1998);

"Measures that are highly correlated (either positively or negatively) are likely
influenced by the same factors, while those that are relatively uncorrelated are likely influenced by
different factors"

What DeCoster is referring to is communality, that is, measures that are common amongst or between factor(s). The underlying aim of Factor Analysis is to use the interaction between these measures as the framework with which to analyze and transform data, reducing dimensionality.

Factor Analysis, while widely used and accepted has been criticized by some due to its misuse. The influence of computing capabilities on the use of Factor Analysis is pinpointed as being to the detriment of 'the underlying theory and theory or methodology of factor analysis' (Stewart, 1981).

D. Variable Analysis & Correalation

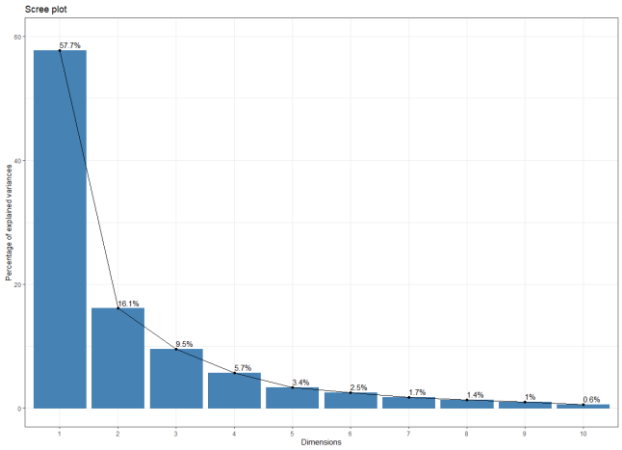
Eigenvalues are the factors by which eigenvectors are stretched. That is to say, they 'point the way' when data is being transformed.

The following Eigenvalues were achieved by conducting initial principal component analysis on the dataframe:

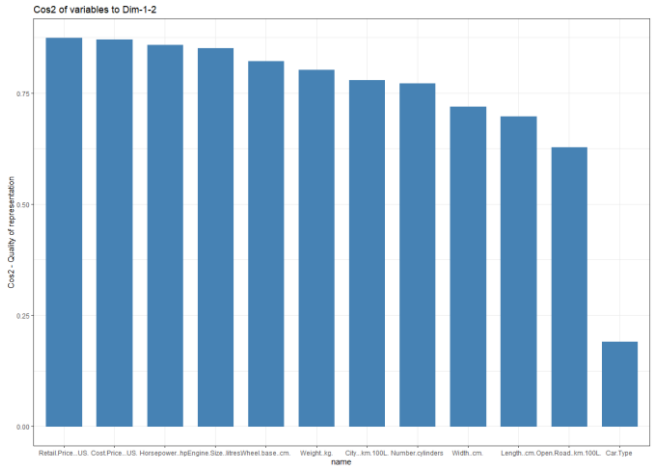
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	6.9259709219	57.716424349	57.71642
Dim.2	1.9353439066	16.127865888	73.84429
Dim.3	1.1432511908	9.527093257	83.37138
Dim.4	0.6831431998	5.692859999	89.06424
Dim.5	0.4057754886	3.381462405	92.44571
Dim.6	0.3016619862	2.513849885	94.95956
Dim.7	0.2098806080	1.749005067	96.70856
Dim.8	0.1635292614	1.362743845	98.07130
Dim.9	0.1194487158	0.995405965	99.06671
Dim.10	0.0761348335	0.634456946	99.70117
Dim.11	0.0350554323	0.292128603	99.99330
Dim.12	0.0008044549	0.006703791	100.00000

As a general rule of thumb, eigenvalues are used as an indication of how many factors to retain. Eigenvalues that are said to be over 1, are generally kept as factors. In this sense, Dimensions 1,2 and 3 would be considered factors.

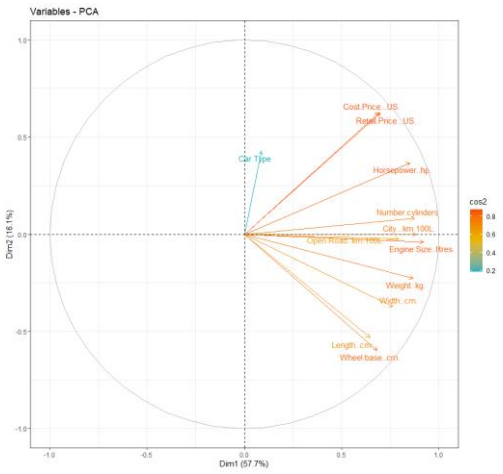
Illustrating this data, we can see two dimensions explain 73.8% of all variances:



Selecting these two dimensions, we can view the quality of representations on our factor map by viewing the cos2 (square cosine, squared coordinates). A high cos2 indicates a good representation of the variable on the principal component. First however, it is important to view the quality of representation on a hierarchical basis:

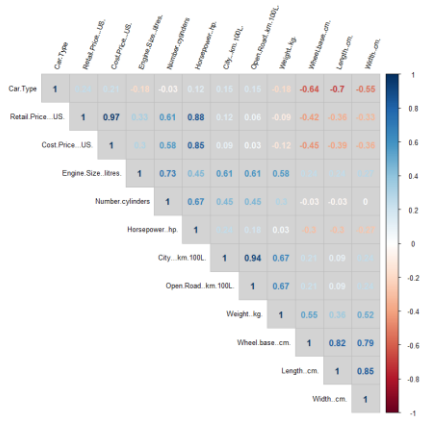


When we cast these variables onto our factor map by cos2, we get the following:

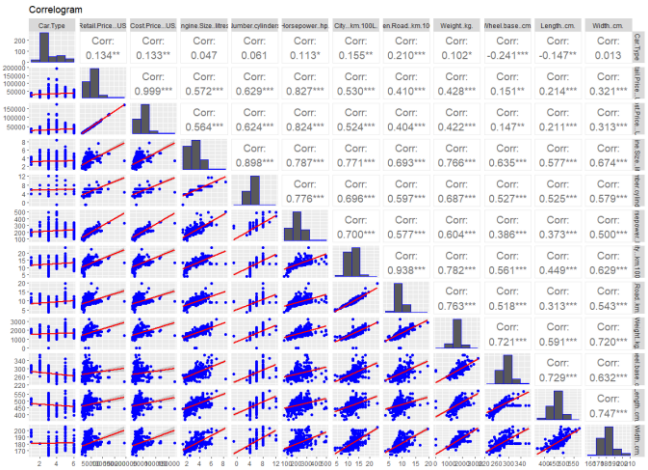


In this case the Car Type variable is positioned close to the circumference of the correlation circle.

As can be seen below, several variables are highly correlated. It is not surprising to see that Horsepower is highly correlated with Retail Price. Engine Size is moderately correlated with weight while the width of the car is highly correlated with its length which insinuates cars with a small width can be said to have a small length.

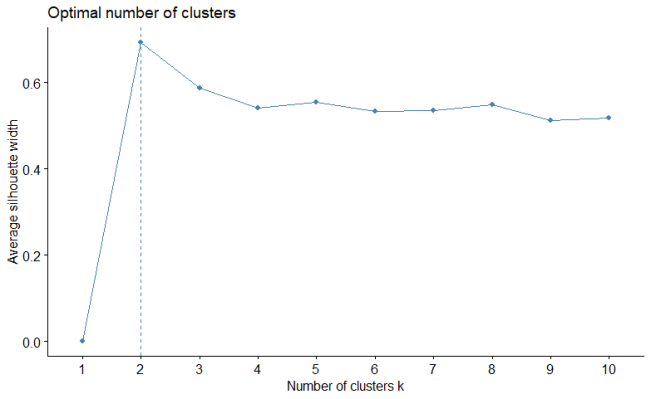


The following correlogram was created so as to show the distribution of data in each variable and map a line of best fit to illustrate the relationship between variables. The more correlated a pair, the more perpendicular this line will be. The observations noticed in the Correlation Plot above, and Correlation Matrix overleaf match our observations from the earlier Correlation Circle.

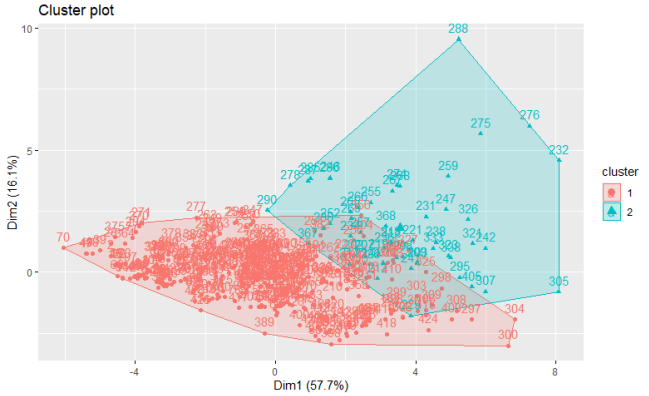


E. Clustering

Using K-Means so as to identify groups which have not been explicitly labelled in the data, we find that only two meaningful clusters exist. By using the silhouette method, to measure the quality of the clustering (how well each object lies within its cluster), the following results were achieved:



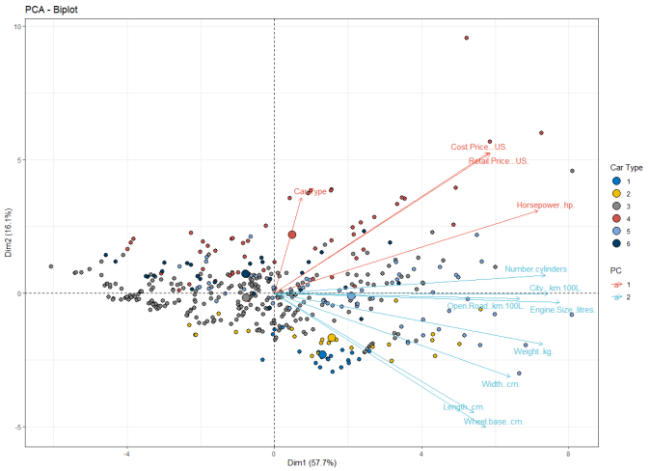
When we graph the suggested number of clusters (2) as below, we get the following results:



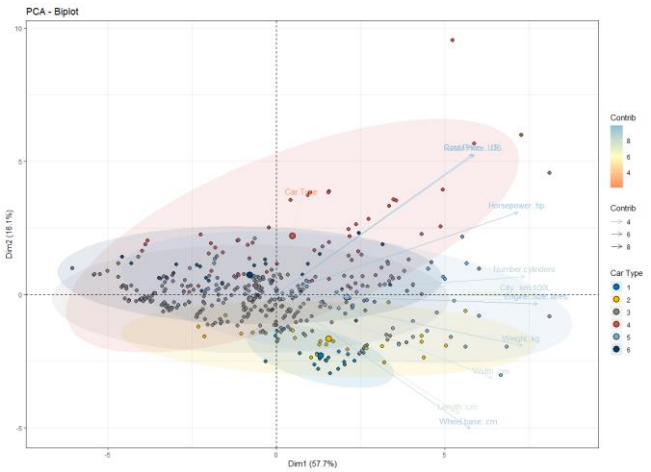
The above graph displays some overlap between the two clusters. Most of the data seems to fall into cluster 1.

F. PCA

Based on eigenvalues and results of k-means clustering, two Principal Components (PC) were chosen. Below we can see the fit of the two PC's:



When we colour the by the target variable (Car Type) and as a measure of the contribution of variables to the specified dimensions of our two clusters, we get the following results:



Car Type 3 appears most centralized, while Car Types 5 and 6 are farthest from the center. Car Group 3 has the largest contribution towards our dataframe.

The following co-ordinates were ascertained for the levels of Car Type. The co-ordinates for a given group are calculated as the mean co-ordinates of the individuals in the group:

CarType	Dim.1	Dim.2
<fct>	<dbl>	<dbl>
1	1.31	-2.30
2	1.55	-1.66
3	-0.771	-0.157
4	0.480	2.21
5	2.09	-0.0928
6	-0.784	0.727

As can be seen, there are a greater number of negative mean co-ordinates for dimension 2 than dimension 1.

The most significantly associated variables with both principal components are set out below:

<u>Dimension 1:</u>

	correlation
Engine.Size..litres.	0.9211930
City...km.100L.	0.8828394
Number.cylinders	0.8745307
Weight..kg.	0.8665406
Horsepower..hp.	0.8507783
Open.Road..km.100L.	0.7919092
Width..cm.	0.7620247
Retail.Price...US.	0.6970403
Cost.Price...US.	0.6915869
Wheel.base..cm.	0.6818395
Length..cm.	0.6433628

Dimension 2

	correlation
Cost.Price...US.	0.6263570
Retail.Price...US.	0.6230148
Car.Type	0.4267701
Horsepower..hp.	0.3668758
Weight..kg.	-0.2270192
Width..cm.	-0.3720579
Length..cm.	-0.5318529
Wheel.base..cm.	-0.5970775

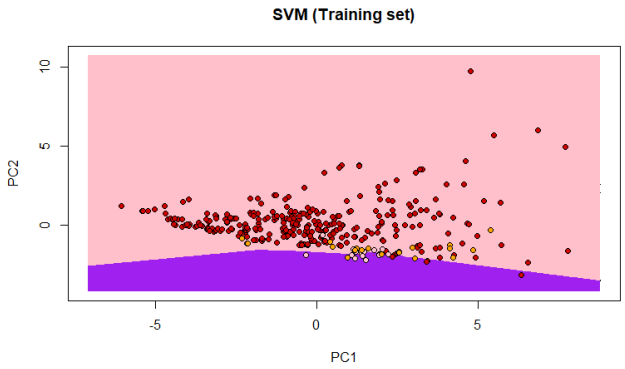
A PCA model was subsequently developed for testing with the dataset split into test and training sets. An 80:20 split was used for conducting PCA. As there are more than two classes, the support vector machine model was utilized. Building on earlier observations, I elected to use two Principal Components.

The following confusion matrix was generated:

	predicted					
actual	1	2	3	4	5	6
1	1	0	3	0	0	0
2	1	0	4	0	0	0
3	0	0	45	3	1	0
4	0	0	7	3	0	0
5	0	0	11	0	1	0
6	0	0	5	1	0	0

50 out of 86 of our test observations were correctly classifed, representing 58% of all observations.

Once trained, the results were called resulting in the following:



As can be seen above, most data pertains to the first principal component This matches our earlier observations in section D. The inference here is that one could reliably remove all data outside of PC1 to leave that which is significant; that which matters.

In summary, three eigenvector values (dimensions) were of significance (<1), of which two were found to be of most significance (73.8% of total variances explained), and thus leading to the discounting of eight other dimensions. K-Means clustering identified two clusters as being most suited. The Principal Component Analysis Model developed had an accuracy rate of 58% and insinuated the data could at the very least be reliably broken down into two factors, if not just one.

II. ANOVA

A. *Introduction*

For this exercise, a dataset was manually created such that it satisfied the condition of normality. The dataset contains three columns, one for sales, another for distribution channel and a third for the type of product sold. There are two distribution channels, and three products. The aim here was to analyse the variance between the distribution channel used and products sold to determine if there is a statistically significant difference between the means of these groups, in which case Tukeys HSD method was to be used to identify where such differences occurred.

B. *Pre-processing*

The following code was used to create the dataframe:

```
#Creating Sales
Sales <- round(runif(1000, min=5, max=100),0)

#Creating Products
Product <- c("Nebuliser", "Inhaler", "CPAP")
Product<- sample(rep(Product, c(333,333,334)))

#Creating Distribution Channel
Channel <- c("Pharmacy", "Direct")
Channel <- sample(rep(Channel,c(500,500)))

#Creating Data Frame

df1 <- data.frame(Sales,Product,Channel)
```

C. *Background*

Ronald Fischer was the founder of ANOVA as we have come to know it. His book, *Statistical Methods for Research Workers* (1925) brought the term ‘variance’ to the fore of statistical analysis and highlighted a new way of analyzing normally distributed data amongst groups.

ANOVA is used to analyze the difference among means. There are several ways in which ANOVA can be used, for example, one-way ANOVA, two-way ANOVA and a fork known as MANOVA for multivariate analysis. ANOVA tests for results of significance between the means of two or more groups to determine if they are statistically significant from one another. ANOVA can be thought of “as an extension of the t-test for two independent samples to more than two groups” (Ostertagová & Ostertag, 2013).

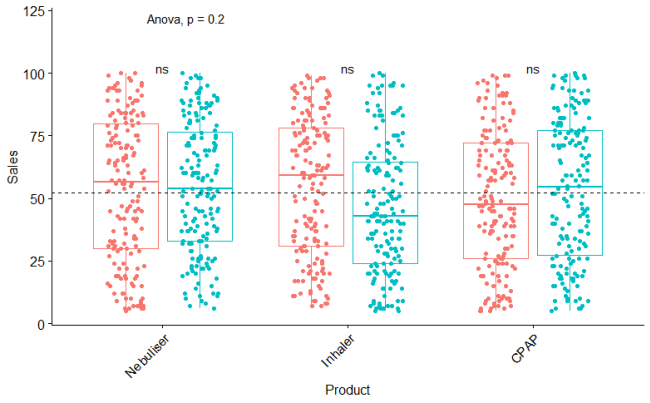
The most common application of ANOVA is in Medical Research through the use of randomized trials, or blind trials (Romaniuk, 2010). In this sense, ANOVA would look to analyze whether the results of one response group for a trial significant differ from the results of another response group, a one-way ANOVA. If the trial was related to people who used both medication A and B and were either Diabetic, or non-Diabetic as a control group, then the test would be to see if the variance between means of these groups were different. This is the two-way equivalent of the aforementioned one-way anova.

Under the hood, an ANOVA test seeks to detect the sources of variation in the values of a dependent variable and divide the total variability into components associated with each source.

The total variability is the sum of squared deviations of each measurement from the overall mean and can be realized as a sum of squares (SS) due to suspected sources of variation (model sum of squares) and a sum of squares (SS) resulting from the error.

D. Analysis

A boxplot was created to visualize the distribution of data within the Products table by Channel:



As can be seen above, it would seem there a distinct different in the distribution of values in the Inhaler Product category between Channels.

Using the Levene Test to test for Homogeneity of Variance, we get the following results:

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   5  1.4381 0.2079
      994
```

As P > 0.05, we accept the null hypothesis. That is; Equal variances exist between our Products and Category variables.

As we seek to conduct the ANOVA test with replication, the following formula was used:

*Sales ~ Product*Channel*

Using this formula, the following results were achieved:

```
                 Df Sum Sq Mean Sq F value  Pr(>F)
Product           2   2405    1202   1.611 0.20017
Channel           1    417     417   0.559 0.45479
Product:Channel   2   7468    3734   5.003 0.00689 **
Residuals       994 741809     746
```

The result confirms there is a statistically significant difference in the variation of observations between the product and channel column.

As such, Tukeys HSD test was used as a post hoc test. Tukeys HSD test helps confirm whether there's a strong chance that an observed numerical change in one value is causally related to an observed change in another value. In order to do this, Tukeys HSD test focus' on the largest value of the difference between two group means. Picking the largest pairwise difference in means allows for controlling of the experiment-wise error rate for all possible pairwise contrasts.

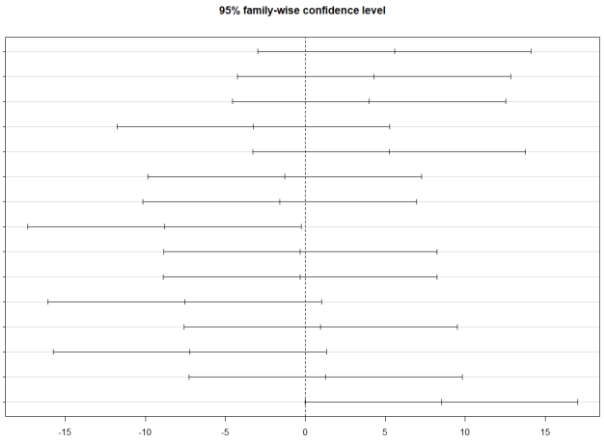
Using this test provided the following results:

```
$Product
                      diff       lwr      upr     p adj
Inhaler-CPAP      -0.8335042 -5.799139 4.132130 0.9180098
Nebuliser-CPAP     2.7941235 -2.171511 7.759758 0.3839062
Nebuliser-Inhaler  3.6276276 -1.341728 8.596983 0.2006011

$Channel
                     diff       lwr     upr     p adj
Pharmacy-Direct -1.291921 -4.682392 2.09855 0.4547911

$`Product:Channel`
                                              diff          lwr         upr     p adj
Inhaler:Direct-CPAP:Direct                5.5718589  -2.96437907 14.1080968 0.4253331
Nebuliser:Direct-CPAP:Direct              4.2947504  -4.24148751 12.8309884 0.7046266
CPAP:Pharmacy-CPAP:Direct                 3.9814974  -4.55474052 12.5177354 0.7673004
Inhaler:Pharmacy-CPAP:Direct             -3.2547049 -11.77807780  5.2686680 0.8853333
Nebuliser:Pharmacy-CPAP:Direct            5.2482891  -3.27508379 13.7716621 0.4935401
Nebuliser:Direct-Inhaler:Direct          -1.2771084  -9.83886582  7.2846490 0.9982188
CPAP:Pharmacy-Inhaler:Direct             -1.5903614 -10.15211883  6.9713959 0.9949509
Inhaler:Pharmacy-Inhaler:Direct          -8.8265637 -17.37549452 -0.2776330 0.0383656
Nebuliser:Pharmacy-Inhaler:Direct        -0.3235697  -8.87250051  8.2253611 0.9999979
CPAP:Pharmacy-Nebuliser:Direct           -0.3132530  -8.87501040  8.2485044 0.9999983
Inhaler:Pharmacy-Nebuliser:Direct        -7.5494553 -16.09838609  0.9994755 0.1188067
Nebuliser:Pharmacy-Nebuliser:Direct       0.9535387  -7.59539207  9.5024695 0.9995648
Inhaler:Pharmacy-CPAP:Pharmacy           -7.2362023 -15.78513307  1.3127285 0.1513291
Nebuliser:Pharmacy-CPAP:Pharmacy          1.2667917  -7.28213906  9.8157225 0.9982745
Nebuliser:Pharmacy-Inhaler:Pharmacy       8.5029940  -0.03309088 17.0390789 0.0515701
```

When we plot the above output, we get the following:



As can be seen, three ranges are very uneven. These ranges would be those relationships with the lowest p values.

The statistically significant result as highlighted (P value < .05), is between the Pharmacy distribution channel and Direct distribution channel. This means the variance between values across these two channels is statistically significant such that there is an observable discrepancy in the price of Inhalers sold in Pharmacies and those Directly. In practical terms, this means if this data were to be based on a real-life business, the pricing profile of inhalers between these distribution channels should be reviewed. Is the discrepancy in values by design or by co-incidence? If it is the latter, how can this be accounted for?

III. Conclusion

Summary conclusions have been included at the end of each section; however it would be remiss not to comment on what further analysis could be done on these reports. For section 1, could we possibly perform an alternative to the PCA and if so, how would it be compared to our results? My first thoughts go to K Nearest Neighbours, which could classify factors on a similarity measure, for example, Euclidian distance. For the business we analysed in section 2, I would interested in adding more variables and columns, and endeavour to perform multivariate regression analysis so as to determine what Sales Price could one expect, if they bought an Inhaler in a Pharmacy **and** *purchased a warranty*?

I feel that is the beauty of statistics, there is always something else you want to analyse, to figure out, to quantify.

IV. References

DeCoster, Jame. (1998). Overview of Factor Analysis. Available at: https://www.researchgate.net/publication/255620387

Fisher, Ronald A. (1921). "On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample". Metron. 1: 3–32.

Ostertagová, E. and Ostertag, O., 2013. Methodology and Application of One-way ANOVA. American Journal of Mechanical Engineering, 1(7), pp.256-261. Available at: https://www.researchgate.net/publication/259291691_Methodology_and_Application_of_One-way_ANOVA

Romaniuk J, 2010. Useful Tips for Analysis of Variance (ANOVA) in Multicenter, Placebo Controlled Clinical Trials. Paper SP06. Quanticate. Available at: https://www.lexjansen.com/phuse/2010/sp/SP06.pdf

Spearman, Charles (1904). "General intelligence objectively determined and measured". American Journal of Psychology. 15 (2): 201–293. doi:10.2307/1412107. JSTOR 1412107.

Stewart, D. (1981). The Application and Misapplication of Factor Analysis in Marketing Research. Journal of Marketing Research, 18(1), 51-62. doi:10.2307/3151313

Thurstone, Louis (1934). "The Vectors of Mind". The Psychological Review. 41: 1–32. doi:10.1037/h0075959.

Williams, B., Onsman, A. and Brown, T., 2010. Exploratory factor analysis: A five-step guide for novices. Australasian Journal of Paramedicine, 8(3).

V. Appendix

[1] *Source for Data Used in PCA:* https://new.censusatschool.org.nz/resource/multivariate-data-sets/