**Chapter 4**

**Analysis of Results and Discussion**

In this chapter the analysis of results of the quantitative techniques used in this research. This chapter includes: the method of data analysis, discussion and analysis of the online survey, demographics, discussion and analysis of the experiment, discussion and analysis of the observation, discussion and analysis in relation to the literature, discussion and analysis in relation to the hypothesis and research questions, and conclusion.

**4.1 Method of Data Analysis**

In this research, numerous methods of data analysis were used for the prototype, direct observation, and online survey. IBM SPSS is the software which was used, focusing on statistics, and aiding the use of several statistical tools to analyse the data gathered.

**4.2 Discussion and Analysis of the Online Survey**

An online survey has been conducted in the early stages of this research as shown in Appendix B. The primary reason for conducting this survey was in order to collect small pieces of statistical information. Such information in this survey included, demographics, experience in the sector, knowledge about Machine Learning, opinions of Machine Learning, current system applications and if they think this will aid their financial planning and prospective futures. Afterwards, the data collected from this survey was assembled using Google Forms and the extracted data was imported to the mentioned IBM SPSS software.

**4.2.1 Demographics**

In the early stages of this research, it was mentioned that a sample of 53 respondents were required to cover a 10% margin of error of the population of auto dealers. These 53 responses were gathered, which were used to carry out descriptive analysis and statistical analysis. The first section, which is the one after the introduction, for the survey included demographic information including gender, age, and experience in the sector. Before the analysis of this demographic data, the researcher had to clean and filter the collected data. As presented in the table below (Figure 4.1), this was done by translating string values such as male and female to 1s and 2s accordingly. This had to be done to every column within the data for it to be readable by IBM SPSS.

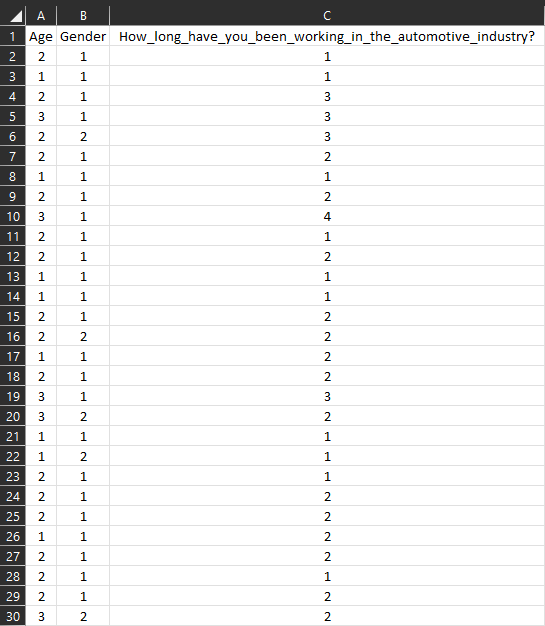


Figure 4.1: Cleaned and Filtered Data

As shown in the chart below (Figure 4.2), shows that from a total of 53 responses there were auto dealers from all age groups. The age group that had the most auto dealers was that of between 30 – 45 having 49.1% (26 auto dealers), followed by that of 15 – 30 years of age carrying 28.3% (15 learners). Combining these two age groups together amounts to 77.4% (41 auto dealers) of the total responses, which shows that most of the auto dealers were on the younger side of adulthood. The rest of the auto dealers were over 45 years of age having 22.6% (12 auto dealers). These figures demonstrated that auto dealers contributed had different age groups, and this was an advantage for this research since various age groups provide diverse opinions.

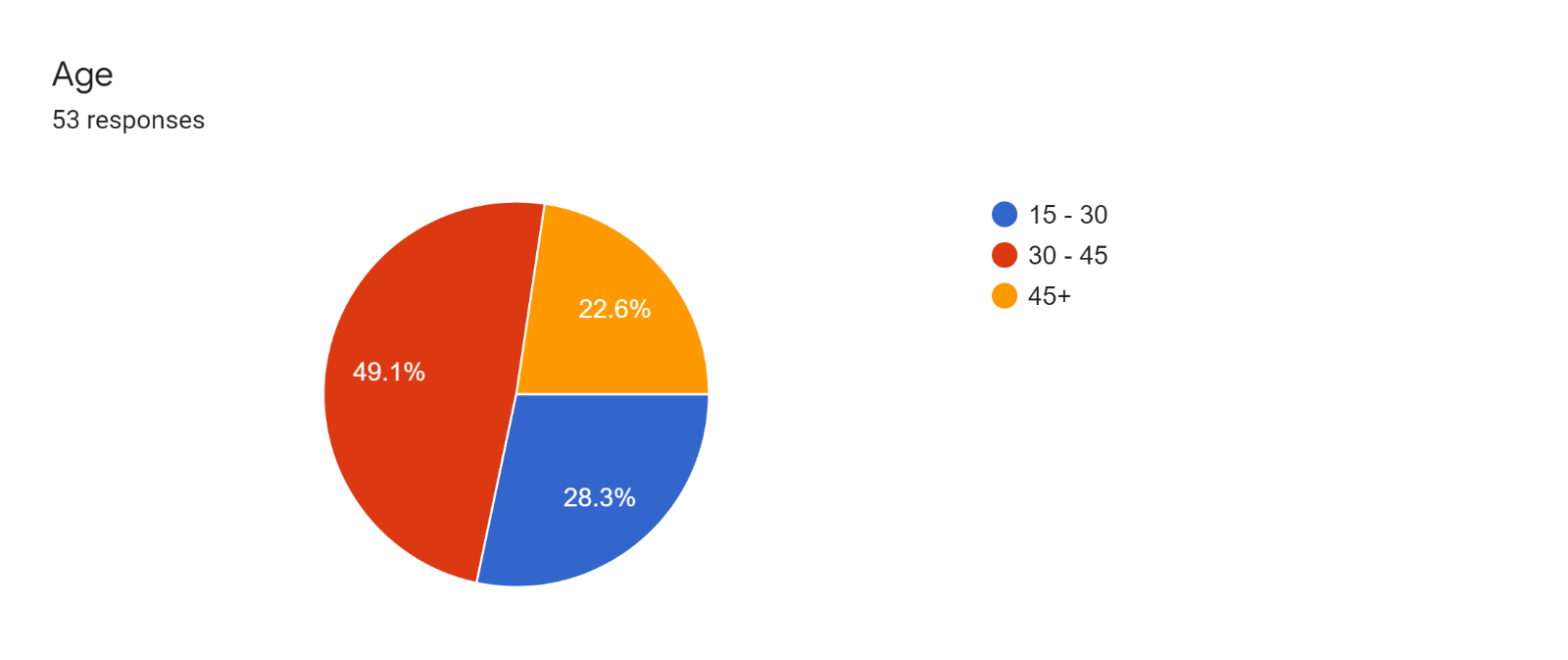


Figure 4.2: Descriptive Statistics Age

As shown in the chart below (Figure 4.3), it can be easily recognized that male auto dealers covered over three quarters of the responses, by having 81.1% (43 auto dealers), followed by female auto dealers carrying 18.9% (10 auto dealers) from the total responses.

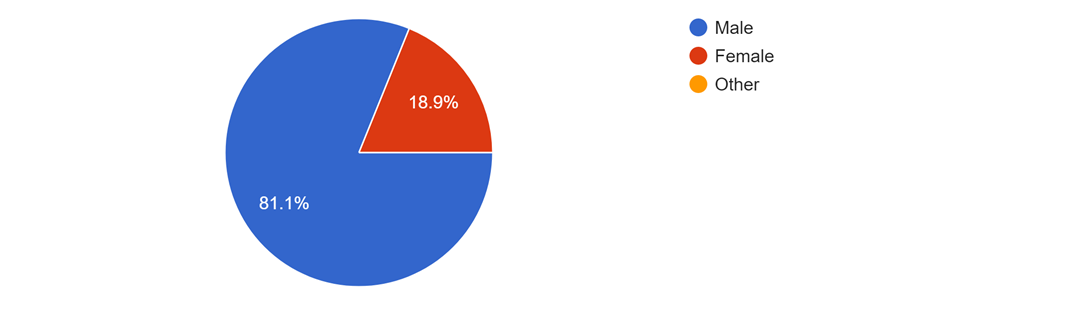


Figure 4.3: Descriptive Statistics Gender

**4.2.2 Chi-Squared Tests**

To analyze the collected data, contrast and compare variables, and cross tabulate; IBM SPSS was used to conduct these analyses. This data analysis method is used to quantitatively examine the correlation between variables.

As shown in Table 4.1 below, the researcher produced the result of the cross tabulation, which shows a statistical analysis between age and whether auto dealers are familiar with Machine Learning. The result shows that there was a higher number of auto dealers familiar with Machine learning which fall in the younger category of 15 - 30, 53.3% of the age group that are familiar with Machine Learning. This is while the older age group category of auto dealers had no respondents which knew about Machine Learning, at 0% all respondents were not aware of Machine Learning. The middle age category result in 18.2% of the age category. This proves that the younger generation are more exposed to technological advancements.

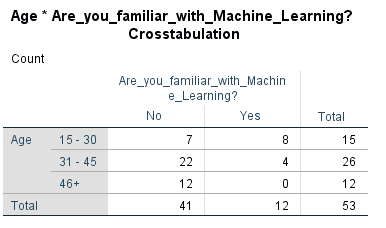


Table 4.1: Cross Tabulation on Age \* Are you familiar with Machine Learning?

In the table below, what the researcher is interested in is the Pearson Chi-Squared, having a p-value of 0.002% and alpha level of 0.05%. This meant that the null hypothesis cannot be accepted, and therefore there is a significant difference between age and whether auto dealers familiar with Machine Learning.

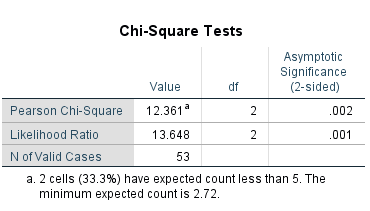


Table 4.2: Chi Squared on Age \* Are you familiar with Machine Learning?

Once conducting cross tabulation, the researcher conducted statistical analysis between age and whether auto dealers thought predictive analytics would be beneficial to their business. Prior to asking this question a small definition was provided to the respondents as shown in Figure B.4.

The first age group, the auto dealers between 15 - 30, have the highest percentage of 100% when it comes to auto dealers that agree that Predictive Analytics could be beneficial to businesses in that age group. The second age group 31 – 45 96.2% of the age category’s respondents agree, while the third age group 91.7% of the age category agree.

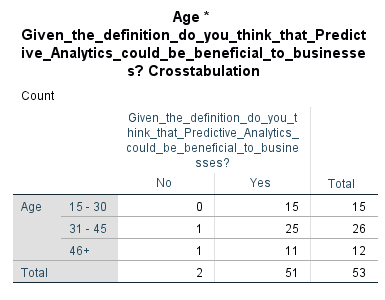


Table 4.3: Cross Tabulation on Age \* Given the definition, do you think that Predictive Analytics could be beneficial to businesses?

In the table below, what the researcher is interested in is the Pearson Chi-Squared, having a p-value of 0.528% and alpha level of 0.05%. This meant that the null hypothesis could be accepted, and therefore there was no significant difference between age and whether auto dealers think Predictive Analytics is beneficial to their business.

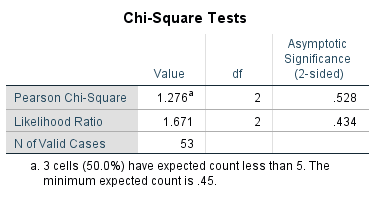


Table 4.4: Chi-Square on Age \* Given the definition, do you think that Predictive Analytics could be beneficial to businesses?

As shown in the table below, the researcher produced the result of the cross tabulation, which shows a statistical analysis between gender and whether auto dealers find it difficult to estimate number of sales to be made per year.

The results show that 69.7% of males and 90.0% of females find it difficult to estimate number of sales per year. In contrast, the total percentage of respondents who find it difficult to estimate the sales made per year are 76.9% males and 23.0% females.

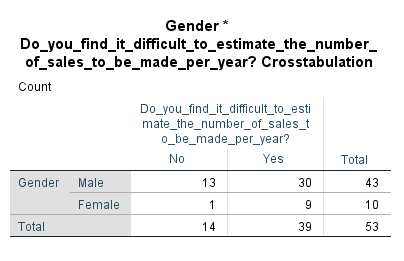


Table 4.5: Cross Tabulation on Gender \* Do you find it difficult to estimate the number of sales to be made per year?

In the table below, the Pearson Chi-Squared have a p-value of 0.191% and alpha level of 0.05%. This meant that the null hypothesis could be accepted, and therefore there was no significant difference between gender and whether respondents find it difficult to estimate the number of sales to be made per year.

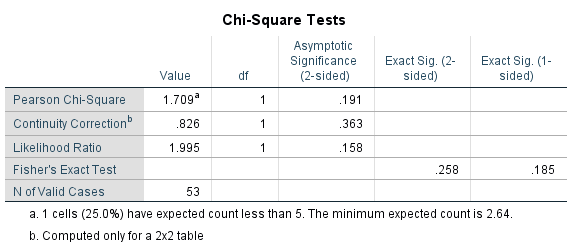


Table 4.6: Chi-Square on Gender \* Do you find it difficult to estimate the number of sales to be made per year?

As shown in the table below, the researcher produced the result of the cross tabulation, which shows a statistical analysis between age and whether auto dealers think that financial planning can be aided with the use of this system.

The results show that the first age group 15 – 30 there is a 100% agreement, the second age group 80.7% of respondents agree, and 50.0% of respondents agree from the third age group. The younger age group shows that they are more open minded to the use of this technology.

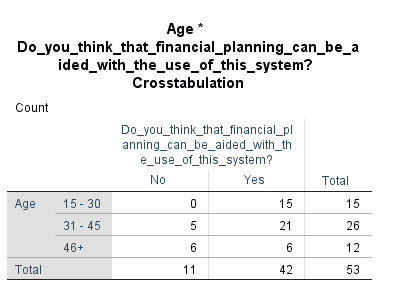


Table 4.7: Cross Tabulation on Age \* Do you think that financial planning can be aided with the use of this system?

In the table below, the Pearson Chi-Squared have a p-value of 0.006% and alpha level of 0.05%. This meant that the null hypothesis cannot be accepted, and therefore there is a significant difference between age and whether auto dealers think that financial planning can be aided with the use of this system.

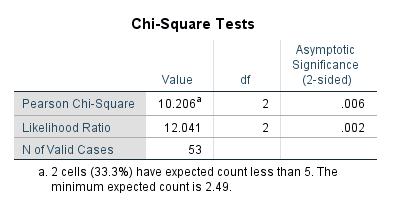


Table 4.8: Chi-Square on Age \* Do you think that financial planning can be aided with the use of this system?

As shown in the table below, the researcher produced the result of the cross tabulation, which shows a statistical analysis between age and whether auto dealers think it is sensible for a business to invest in such a system. The results show that the first age group 15 – 30 is the highest since there is a 93.3% agreement, the second age group 92.3% of respondents agree, and the third age group has 58.3% of respondents in agreement.

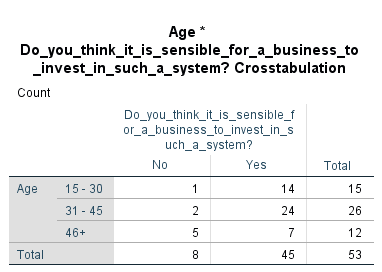


Table 4.9: Cross Tabulation on Age \* Do you think it is sensible for a business to invest in such a system?

In the table below, the Pearson Chi-Squared have a p-value of 0.014% and alpha level of 0.05%. This meant that the null hypothesis cannot be accepted, and therefore there is a significant difference between age and whether auto dealers think it is sensible for a business to invest in such a system.

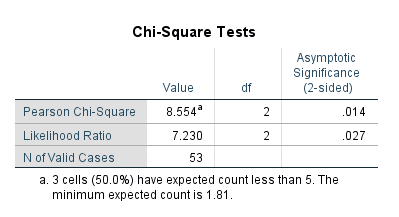
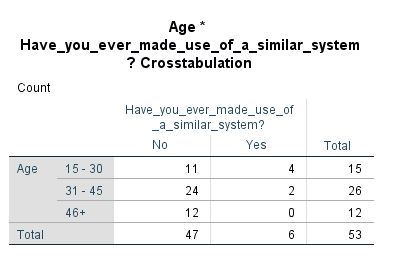


Table 4.10: Chi Squared on Age \* Do you think it is sensible for a business to invest in such a system?

As shown in the table below, the researcher produced the result of the cross tabulation, which shows a statistical analysis between age and whether auto dealers have ever made use of a similar system.

The results show that the first age group 15 – 30 has the highest agreement of 26.7% agreement, the second age group 7.7% of respondents agree, and 0% of respondents agree from the third age group. Attempts have been made by the younger age group in order to try similar prediction systems.

Table 4.11: Cross Tabulation on Age \* Have you ever made use of a similar system?



In the table below, the Pearson Chi-Squared have a p-value of 0.067% and alpha level of 0.05%. This meant that the null hypothesis can be accepted, and therefore there is no significant difference between age and whether auto dealers have ever made use of a similar system.

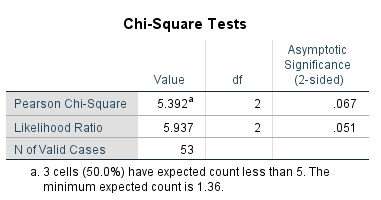


Table 4.12: Chi Squared on Age \* Have you ever made use of a similar system?

As shown in the table below, the researcher produced the result of the cross tabulation, which shows a statistical analysis between gender and whether auto dealers recommend the use of a system which estimated the number of sales with the given specifications.

The results show that 90.6% of males and 80.0% of females find it difficult to estimate number of sales per year. The total percentage of respondents who find it difficult to estimate the sales made per year are 83.0% males and 17.0% females.

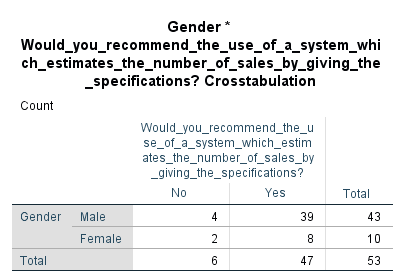


Table 4.13: Cross Tabulation on Gender \* Would you recommend the use of a system which estimates the number of sales by giving the specifications?

In the table below, the Pearson Chi-Squared have a p-value of 0.336% and alpha level of 0.05%. This meant that the null hypothesis can be accepted, and therefore there is no significant difference between gender and whether auto dealers recommend the use of a system which estimated the number of sales with the given specifications

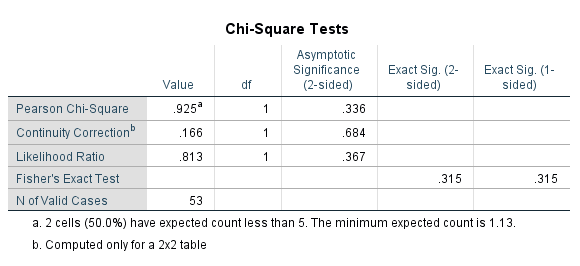


Table 4.14: Chi Squared on Gender \* Would you recommend the use of a system which estimates the number of sales by giving the specifications?

As shown in the table below, the researcher produced the result of the cross tabulation, which shows a statistical analysis between age and whether auto dealers think the shown tools are easily seen and that the features and functions are clearly identified.

The results show that the first age group 15 – 30 has the highest agreement of 100% agreement, the second age group 96.2% of respondents agree, and 83.3% of respondents agree from the third age group. The younger age group finds the current application and its features clearly identified.

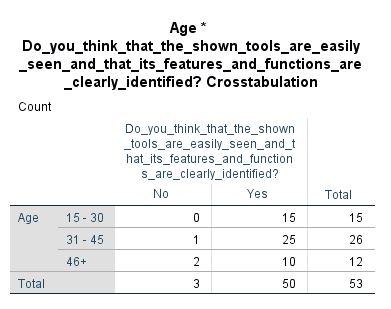


Table 4.15: Cross Tabulation on Age \* Do you think that the shown tools are easily seen and that its features and functions are clearly identified?

In the table below, the Pearson Chi-Squared have a p-value of 0.151% and alpha level of 0.05%. This meant that the null hypothesis can be accepted, and therefore there is no significant difference between age and whether auto dealers think the shown tools are easily seen and that the features and functions are clearly identified.

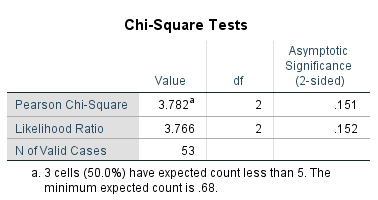


Table 4.16: Chi Squared on Age \* Do you think that the shown tools are easily seen and that its features and functions are clearly identified?

**4.2.3 ANOVA Test**

The researcher also produced another test, the One-Way ANOVA, concerning age and a relevant question. The ANOVA test was used for two questions. The first question: “From what you know until now, do you think Machine Learning is useful in the automotive industry?” having multiple choice options of: “Highly Agree”, “Agree”, “Disagree”, and “Highly Disagree”.

In the table below, the Pearson Chi Squared for the age had a p-value of 0.03% which is smaller than 0.05%. This meant that the null hypothesis cannot be accepted, and therefore there is a significant difference between age and whether auto dealers think Machine Learning is useful in the automotive industry. In contrast, gender had a p-value of 0.646% which is greater than 0.05%. This meant that we can accept the null hypothesis and therefore there is no significant difference between gender and whether auto dealers think Machine Learning is useful in the automotive industry.

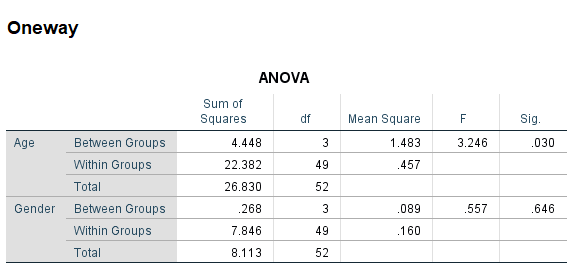


Table 4.17: One-Way ANOVA on Age & Gender \* From what you know until now, do you think Machine Learning is useful in the automotive industry?

In the table below, the Pearson Chi Squared for the age had a p-value of 0.108% which is greater than 0.05%. This meant that the null hypothesis can be accepted, and therefore there is no significant difference between age and whether auto dealers think this system can provide a good sight of what's to come in the near future. In contrast, gender had a p-value of 0.553% which is greater than 0.05%. This meant that we can accept the null hypothesis and therefore there is no significant difference between gender and whether auto dealers think this system can provide a good sight of what's to come in the near future.

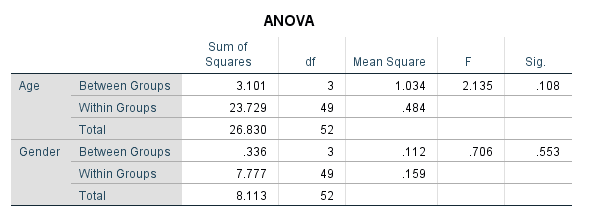


Table 4.18: One-Way ANOVA on Age & Gender \* Do you think this system can provide a good sight of what's to come in the near future?

**4.2.4 T-Test**

The researcher displayed a parallel test to the One-Way ANOVA, the T-Test, concerning age and the question “How long have you been working in the automotive industry?”. The primary distinction between the One-Way ANOVA and the T-Test, is that One-Way ANOVA uses three or more variables, for example age groups 15 - 30, 31 - 45, and 46+, in contrast to the T-Test which makes use of only two variables, such as gender, Male and Female.

For this test, the researcher had to consider the value of the 2-tailed significance, where when its value is greater than 0.05 the null hypothesis is accepted and when the value it less than 0.05 it cannot be accepted. In the table below, the Pearson Chi-Squared has a p-value of 0.132% and alpha level of 0.05%. This meant that the null hypothesis can be accepted, and therefore there was no significant difference between age and how long has the auto dealer been working in the automotive industry.

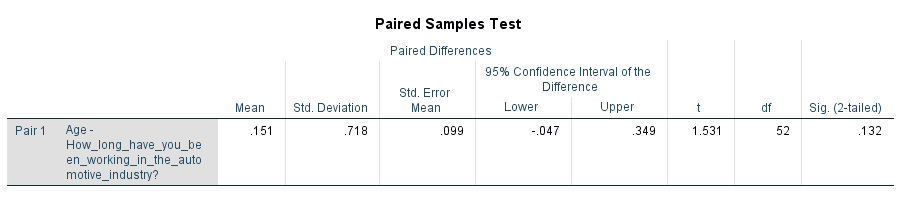


Table 4.19: T-Test on Age \* How long have you been working in the automotive industry?

The table below shows the T-Test between gender and the question “Which Machine Learning technique would you find most suitable for your organization?”. Having the Pearson Chi-Squared with a p-value of 0.000% and alpha level of 0.05%. This meant that the null hypothesis cannot be accepted, and therefore there was a significant difference between gender and the auto dealer choice for machine learning technique would be suitable for their organization.

Table

Description automatically generated

Table 4.20: T-Test on Age \* Which Machine Learning technique would you find most suitable for your organization?

**4.3 Analysis and Discussion in Relation to the Experiment**

The experiment implementation was following via the pipeline mentioned in 3.3.2 in the preceding chapter. Moreover, four scripts were written to output different models for better evaluation. Also, the researcher implemented these scripts by using the original value of the dataset and that same value however in thousands so this can vary among different datasets if needed.

The dataset included monthly domestic auto sales based in the United States of America. Further, the dataset included the thousands of units sold each month, rather than of the sales amount. After the cleaning and resampling of the data, the time series was constructed and decomposed, as seen in figure 4.4.

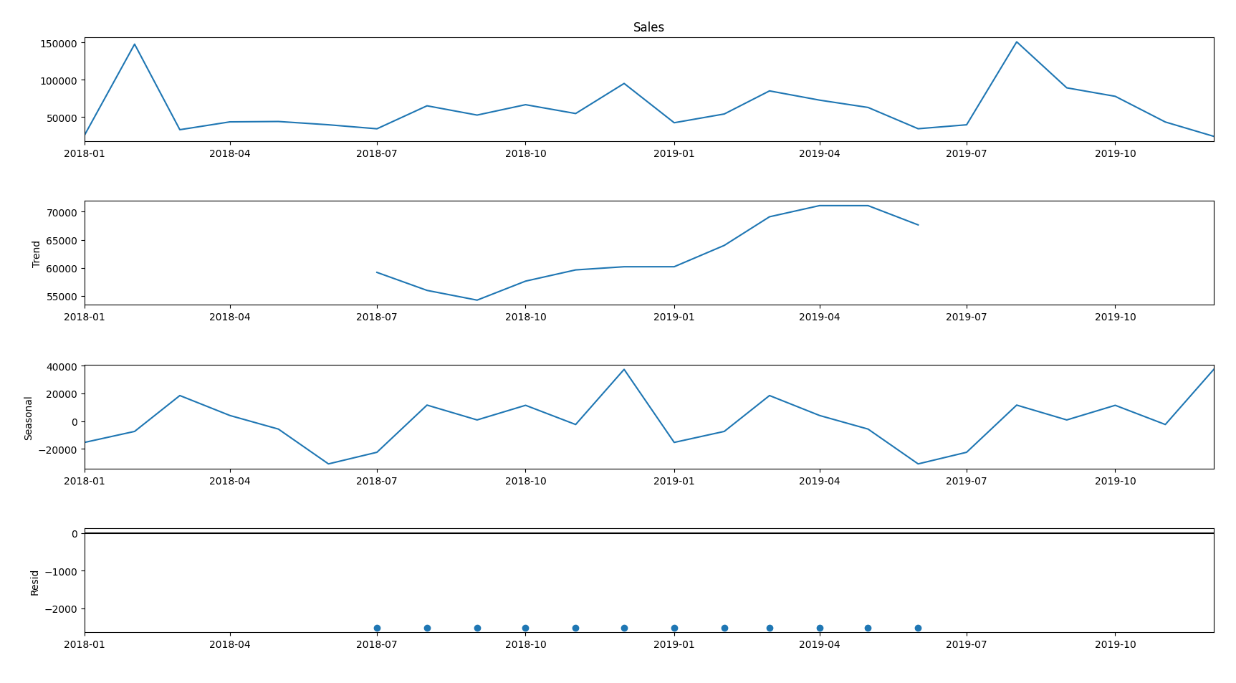


Figure 4.4: Domestic automobile dataset decomposition

The initial line chart exhibits the original time series of one year of sales data whereas the ‘Trend’ chart demonstrates the element of trend of the series, which is turning down towards the time series. In addition, the seasonal element proved to be constant, concluding in the necessity of using a seasonal model. Finally, the residual element was very slightly fluctuating throughout the time series.

Subsequently, the researcher utilized the grid-search method on the time series to determine the optimum parameters for (p,d,q) and (P,D,Q), and make use of them within the Seasonal-ARIMA model. Resultantly, the grid-search technique returned different parameter sequences with distinctive Akaike's Information Criterion (AIC) scores, having the lowest AIC score combination selected. In this case, ARIMA(0,1,1)(1,1,0)[12] was recommended by the prototype, holding an AIC score of 269.674. Therefore, the model was created with these parameters and fit to the training set and graphed with the testing set for visualisation purposes.

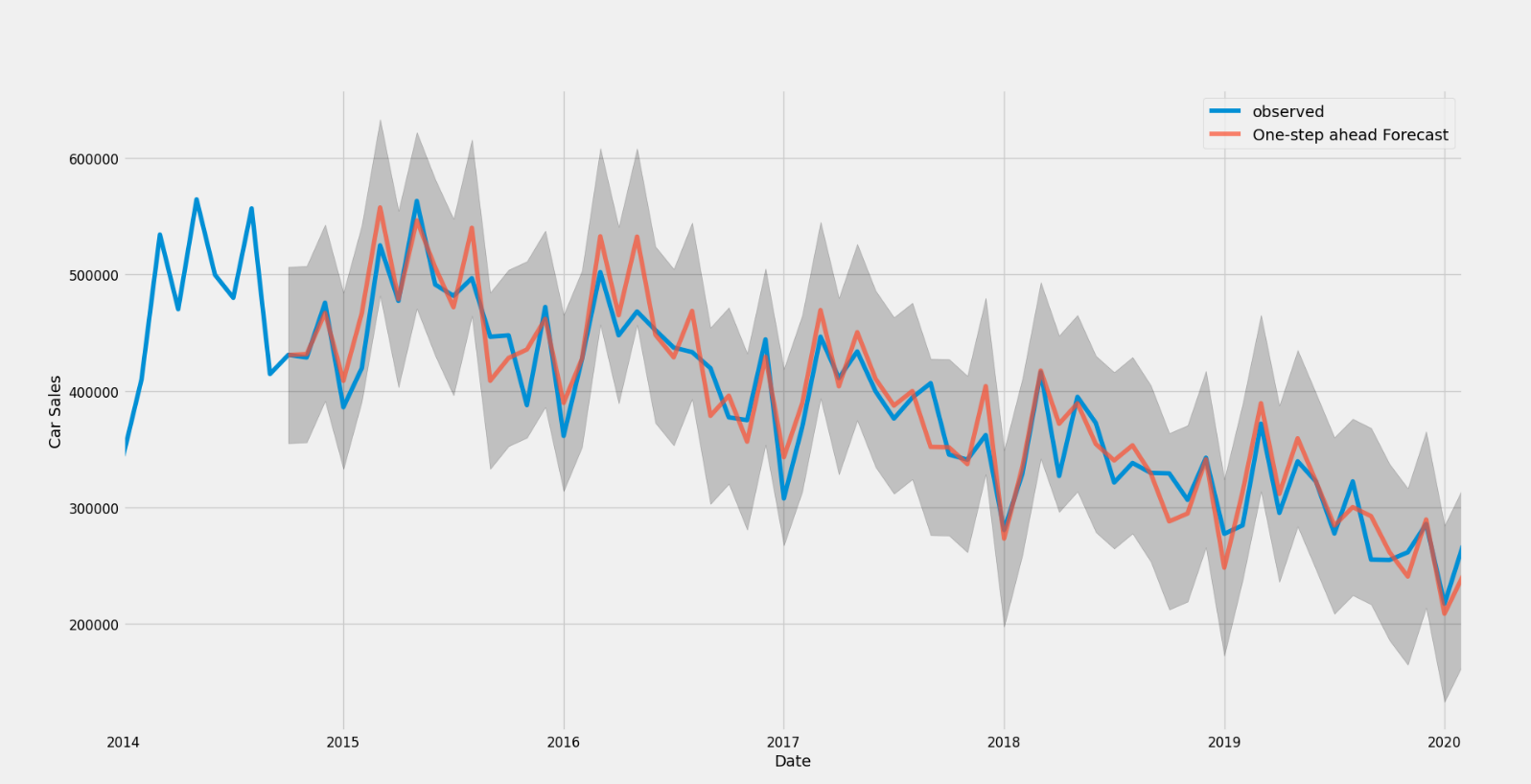


Figure 4.5: Grid-search plot for domestic auto dataset

Figure 4.5 above, displays this line chart, showing that the observed (blue) and forecasted (orange) values are fairly firmly modelled to one another. The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) for the model forecasted values to calculate the error amount and validate the performance of the model. Moreover, when compared to the real values, the forecasts are considered to be relatively close. This was condemned by the metrics (as seen in table 4.3) as the MSE of 10909.73 and RMSE of 104.45 were fairly low when compared to the minimum (200,000) and maximum (564,577) values in the data.

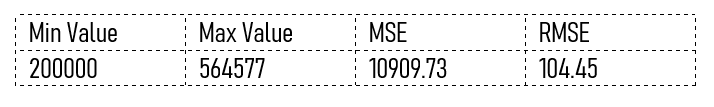


Table 4.3: Grid-search metrics for domestic auto dataset

The forecasted values for 36 months in the future were then produced by the prototype, whilst using the model. Furthermore, these values were marked on the same chart as the original dataset values to grasp a clearer understanding of future sales in contrast to current data. Figure 4.6 displays this chart and as can be displayed, the model captured the downward trend as well as the seasonal aspect of the data. As a result, it can be said that with the low (RMSE and MSE) metrics and visualisation, the performance of the model is very good.

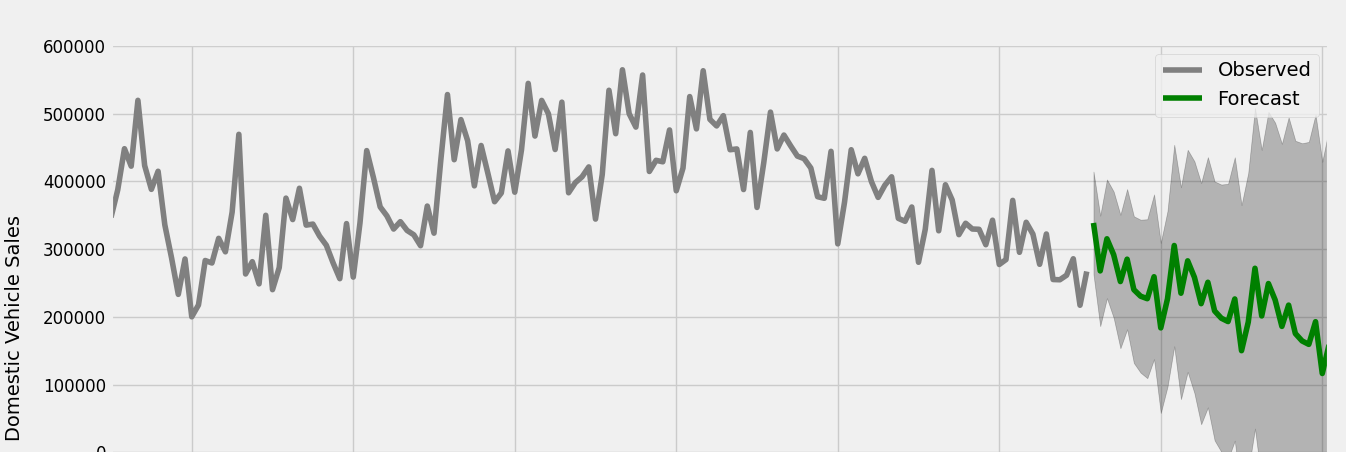
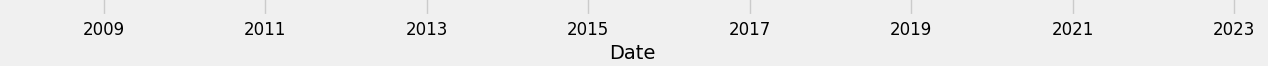


Figure 4.6: Grid-search future forecast for domestic auto dataset

With the help of these experiments, the researcher concluded that the grid-search method proved to be accurate. Moreover, this was shown in the AIC, MSE, RMSE and visualisation plots in the output of the prototype. The best model performance was presented by the grid-search method on the Domestic Auto Sales data outputting a low RMSE metric whilst capturing trend and seasonality elements.

**4.4 Analysis and Discussion in Relation to the Literature**

The similarity with the literature review had to be conducted after the analysis of results for both the online survey and the experiment. What Mamgain et al argued is that any missing values within the data set are handled by omitting the entire row which contains a missing value, while also omitting the missing data. This has been tested within this experiment as the question “Which Machine Learning technique would you find most suitable for your organization?” had a total of 11 responses and was the only question which was not compulsory. The analysis was first conducted with the null values which resulted in problems with the outcome and required for the null values to be provided a value or be ommitted from the analysis completely. This has proved what Mamgain et al argued to be true.

Schrider and Kern mentioned the benefits of using supervised machine learning and in addition to accurate prediction, one of the main benefits of using supervised machine learning is the ability to circumvent using ideal, parametric simulations of the data when labeled training data can be attained from practical observation. Similarly, the experiment in this research provided the ability to avoid inaccurate estimations and practical observation from the dataset gathered was used in order to create predictions via labeled training. They also argued that Supervised Machine Learning can be trained to identify occurences as they are in nature, which can be interpreted accurately with this experiment, since it focuses on the concept that the economy as well as the business will flow in the same manner as it was flowing and there are no supernatural occurences which changed the flow of any of these variables, such as covid-19 as a recent example.

According to the SAS, the expected outcome of this approach is to surpass what has happened to provide the best evaluation of what the impending outcome shall be. This also comes to an agreement with Dairu and Shilong, such a system once tested, and accuracy is confirmed, has reduced uncertainty while also anticipating change in the market, affecting investors, shareholders and customers and if the use of such a product were in real time by simply inputting data from past observations can create a more comforting prediction to the business to reduce preceding anticipations while also remaining accurate.

**4.5 Analysis and Discussion in Relation to the Hypothesis and Research Questions**

For the hypothesis to be examined, the researcher required to perform discussion and analysis of results. This was accomplished by using a working solution, while also having the research questions being answered. Stated earlier in this study, the hypothesis of this research was “Supervised machine learning, can predict auto dealer car sales”. As variables, this hypothesis has Supervised Machine Learning as the direct variable, and Auto Dealer Car Sales as the indirect variable. The implemented solution covered the direct variable, while the online survey and direct observation were implemented to cover the indirect variable. This emphasized whether a Supervised Machine Learning technique can truly predict the forecast of Auto Dealer Car Sales for businesses in Malta which attempt to make use of these techniques.

It was noted in multiple questions that the age group affects greatly whether auto dealers would be interested in using machine learning in their business to predict an outcome, Figures 4.9 and 4.10 that the older audience proved to be more confident in more traditional ways of selling cars in their business and would not be interested in such new technology. In contrast, there was much less problems of acceptance and understanding with the younger audience when explaining the function of this experiment and its use. Figures 4.13 and 4.14, however, show that the 88.7% of respondents would recommend the use of such a system when given the specifications, meaning that although the concept was harder to grasp for the older groups, a vast majority of the respondents would still attempt the use of such a product. Figures 4.15, 4.16, 4.18 show that the prediction of the business with the correct data is possible and recommended from the auto dealers. Therefore, these prove this hypothesis to be true, meaning this system of a Supervised Machine Learning Technique can predict Auto Dealer Sales.

Two research questions had to be tackled, where the first research question was, “Does applying supervised machine learning technique increases auto dealer car sales?”, the second research question, “Can machine learning using a supervised learning method accurately predict auto dealer sales?”.

Reasearch question 1 was tackled through the prototype in conjunction with the conducted online survey, auto dealers have communicated their anonymous opinions and few which have used have seen minor increases in sales. To tackle Research question 2, a python-based supervised machine learning program was developed.

Figures 4.16 and 4.18 showed the average ratings of filtered content, content variability and relevant training to be quite high, so this demonstrates that these factors are a good solution to predict Auto Dealer Sales. Table 4.9 and 4.10 showed that 93.3% from the 15 – 30 age category as well as 92.3% from the 31 – 45 age category were in agreement to the use of the experiment to aid their business. In contrast, the age group 46+ were in 58.3% in agreement which shows a different perspective in older age group. All the questions proved that age was the main factor on whether the system shown is used to affect a business and gender had no further affects.

**4.6 Conclusion**

In this chapter, the analysis of the results of this research was discussed, including: the method of data analysis, discussion and analysis of the online survey, demographics, discussion and analysis of the experiment, discussion and analysis of the observation, discussion and analysis in relation to the literature, discussion and analysis in relation to the hypothesis and research questions. In the next and final chapter, a conclusion and recommendations will be outlined. This chapter will include: the limitations, and future work.

**Appendix B**

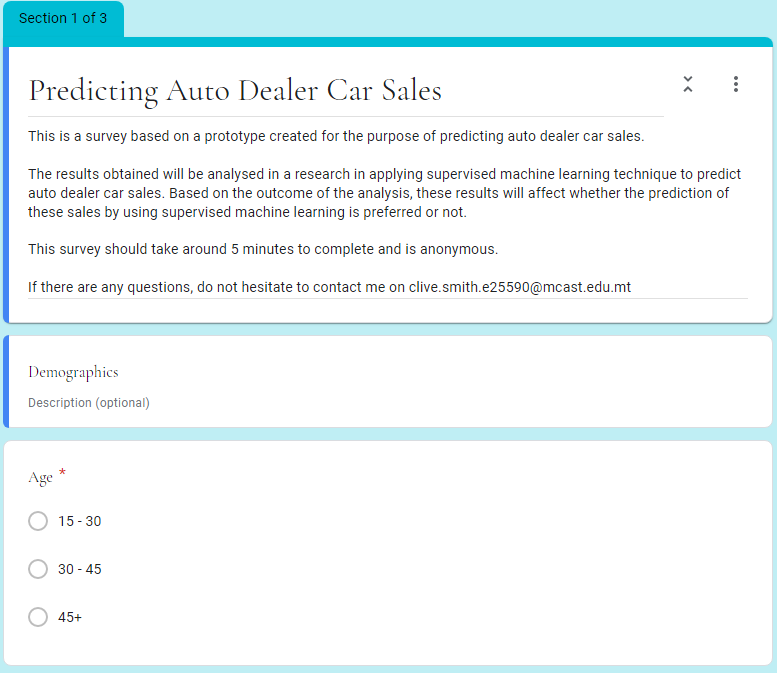


Figure B.1: Online Survey Introduction and Initial Demographics

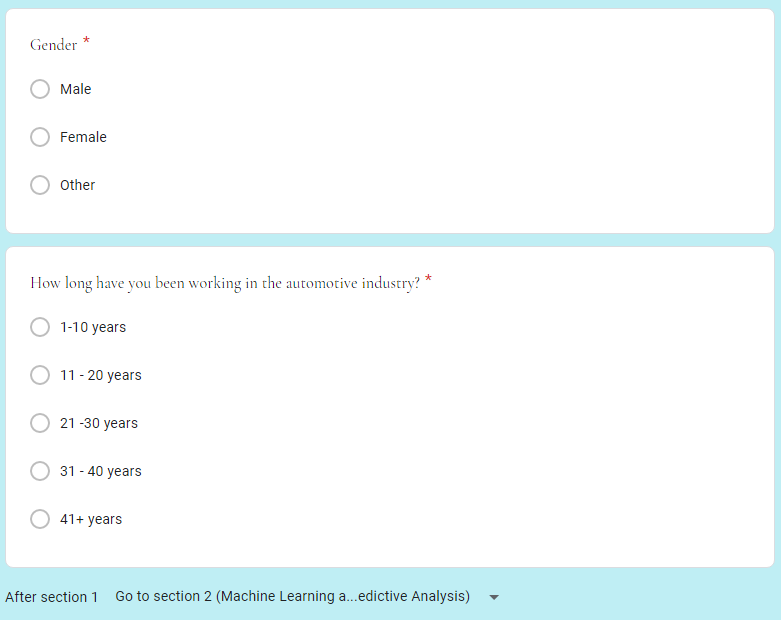


Figure B.2: Online SurveyContinuation of Demographic Information

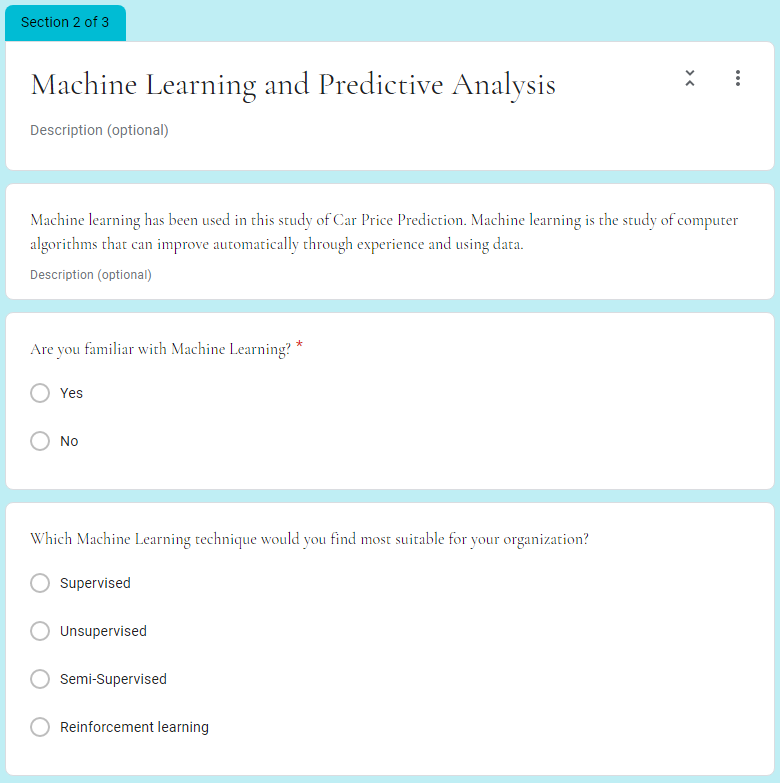


Figure B.3: Online Survey Machine Learning Definition and Knowledge Part 1

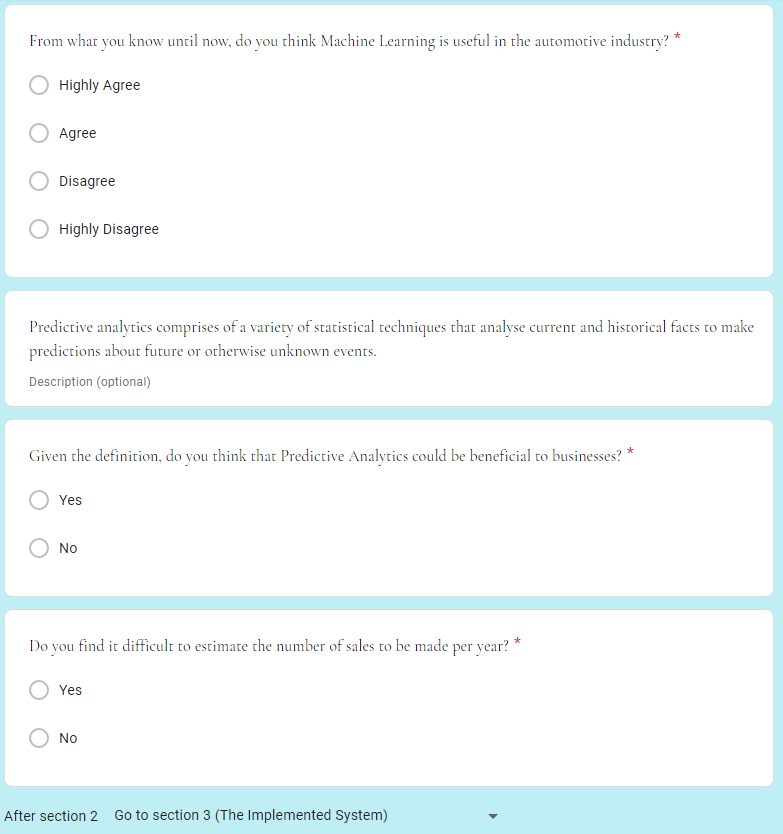


Figure B.4: Online Survey Machine Learning Knowledge Part 2 and Predictive Analytics Definition and Knowledge

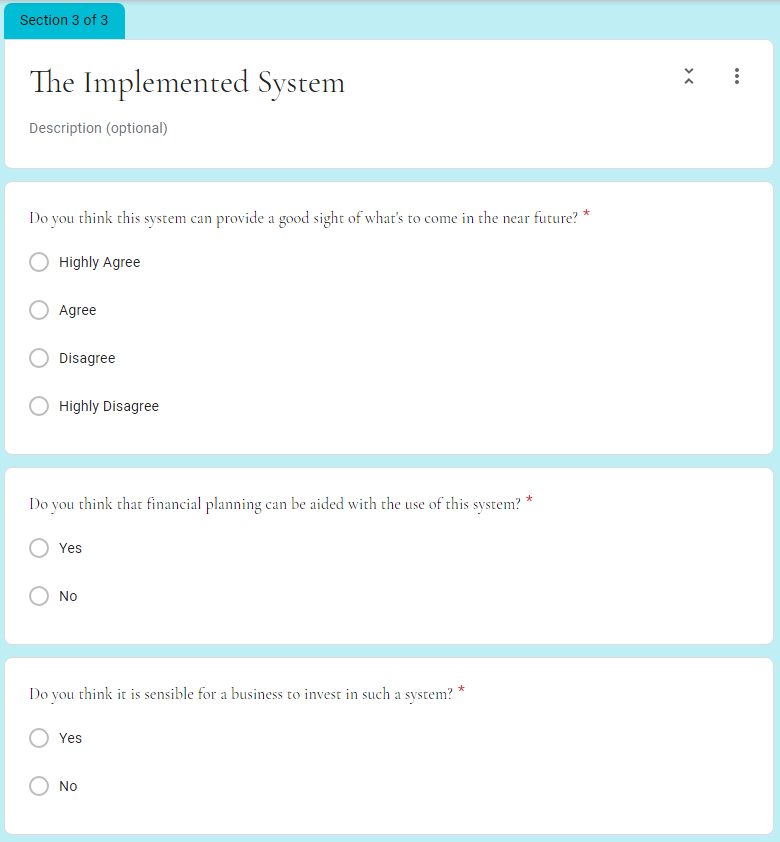


Figure B.5: Online Survey Current Implemented System Part 1

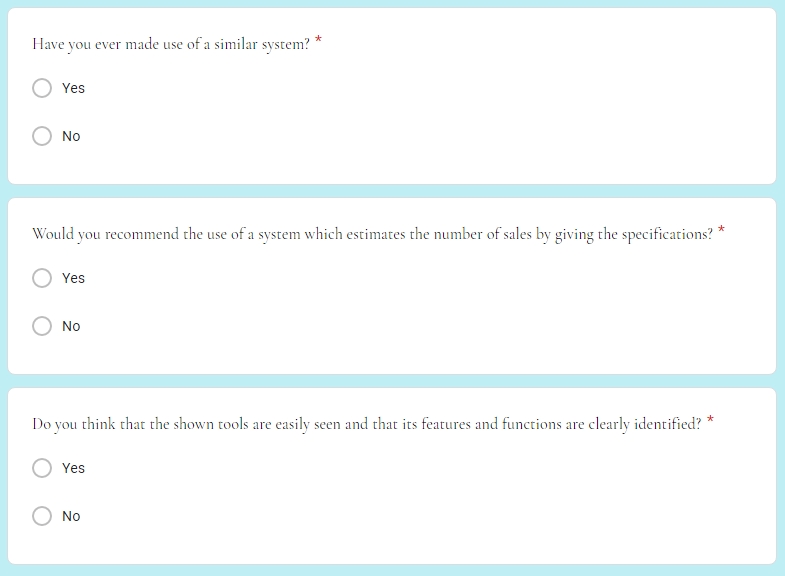


Figure B.6: Online Survey Current Implemented System Part 2