Task 1: Comments to Interview Notes

i. 80% non-response rate is normal, especially in online surveys. Using extrapolation in linear regression is unreliable and not accurate. First of all, the regression is performed with existing data of 2 variables and the regression line is used as a framework for data analysis on the existing data. Extrapolating this relationship beyond the existing data range can yield unpredictable and/or unrealistic outcomes. Farsan: A proper method of fixing the non-response is to ask again in a different time/situation to decrease the non-response rate.

ii. Panels or focus groups is a good way of collecting anecdotal data (not quantitative). Also, we would extend the age upper limit to 70 or 80. Farsan: The groups however can have influence on each other’s minds when talking in groups.

iii. Providing incentives is another good way to increase the response rate. Farsan: However it can affect the quality of data by creating the rate of survey responses that are not genuine only to increase their chances of winning the prize.

iv. A good way to protect customers’ personal info privacy and security (Carl, can you take a look at this?)

Farsan: IP address is an identifying piece of information. In order to obfuscate the identity, IP addresses should be replaced by the region the IP address is for. The region’s size should be defined to compensate between it being smaller size that pinpoints user’s precise location and larger size which results in inaccuracy of the information.

v. Carl

Farsan: We can’t ensure the privacy by keeping the information in two files alone, although it’s a good start. We have to have a linking parameter to connect the IP address to the information for later relinking the parameters and that would act as user id. Another measure of privacy we can build in order to protect user’s information is to encrypt the files or to encrypt each pieces of information and at the same time control the access information to the file by restricting the access to privileged and controller users.

vi. It is not ethical to make it mandatory for customers to provide their income info or SIN number when collecting data. A better approach is to give customers to choose income levels in different ranges, instead of providing exact dollar amounts. If a customer still does not want to provide his/her income level, N/A should be an option. Farsan: Also putting a mandatory parameter like income in the survey increases the non-response rate due to the customers not willing to share this information.

vii. We do not see the point of obtaining SIN numbers if customers refuse to provide their income level. Also, keeping SIN numbers is a security and privacy burden for Sleepy. Farsan: The same points for vi are also valid here. Any mandatory field which jeopardize customers privacy and personal information would increase the non-responsivity of the servey.

viii. Identifying multiple competitors and coding them with values is a good way to build a multiple regression model for statistical analysis. Farsan: A better way to gather information that we may not be aware of is to add a text field for the customer to write the name of the vendor for the “Rest” case.

xiv. Farsan: Increasing of the prices is another event that needs to be added to the system. It creates two separate data sets that needs to be analysed individually to increase the accuracy of the analysis.

x.Obtaining confidential data from a competitor and using it to one’s own advantage is strictly unethical and even illegal. The data scientist should only use data obtained from legal means and keep it private and secure. Farsan: Also obtaining information from an individual would not ensure the accuracy and validity of the data and it may cause invalid and misinterpreted results.

xi. Before choosing multiple regression, we need to verify if the model is normally distributed by using various graphical techniques. Some useful ones include normal probability plot, absolute values of residuals against fitted values, residuals in order of their data collection, and residuals against each predictor variable plot. Farsan: Also logistic regression and multiple regression models have different use cases. Logistic regression is used to classify categorized variables as multiple regression is used for continuous numerical data.

xii. Removing an outlier needs to be treated with caution. Before doing so, the value of the potential outlier needs to be inspected and judgement is then made from a knowledgeable data scientist. Farsan: or a domain expert. The data removed and modified has to be logged so that we can audit the results when validating the model and the results. It is a good practice to also include it in the final report.

xiii. Defining an outlier as 2 standard deviations away from the mean is incorrect. There are several ways of identifying outliers, one of which is by creating a normal probability plot and checking data points that are far away from the straight line.

xiv. Using R-squared and goodness of fit test are good indicators of a regression model. Additionally, with multi variable regression models, R-squared adjusted is better.

Farsan: Also using the training speed versus error diagram when performing cross-validation is another good measure to make sure that overfitting does not happen.

xv. While in this case, a p-value of 0.4 failed to reject the null hypothesis that Barebones was preferred, it is incorrect to say the alternative hypothesis is rejected. Also, null hypothesis statements should be formed as negative statements. For instance, Barebones was not the preferred choice. Therefore, a p-value of 0.4 means there is no evidence to suggest that Barebones was not the preferred choice (i.e. we do not know if Barebones was the preferred choice).

xvi. Decreasing the p-value down to 0.01 is incorrect and against the rule of thumb of 0.05. Additionally, quality problems need to be elaborated.

xvii. Data cleaning is an important step before performing data analysis. We agree with what the scientist did here. Farsan: as long as it is logged and mentioned in the final report. Removing the sample data has to be done with cautions as there might be ways to fix the data instead of removing precious samples.

xviii. Creating a prediction model that is 100% accurate is very likely that there is overfitting of data. Farsan: We need to use cross validation to make sure that the model is not overfitted.

xix. Having the data scientist perform a thorough review is good but we need another person with enough statistical knowledge to review all the procedure, techniques, and the data used to get a different perspective.

xx.It seems like there is a strong bias in the data scientist’s mind before the statisical exercise began. He/She discussed with the marketing team, who is desperate for more budget to continue the campaign. It seems that the data scientist designed the procedure and statistical analysis to confirm the bias and we are not confident enough to trust his/her findings.

Task 2: Comments on Statistical Analysis from Esta

While it is true that there is no relationship between age and preference for Barebones, it is more accurate to state that there is no significant effect of age on the preference of Barebones.

The second sentence in the Data Scientist’s conclusion is incorrect, as one cannot be 95% confident about the relationship or the independence between age and preference of Barebones. We can only be 95% confident that a point estimate will fall into this regression model.

The linear regression line equation of y = -0.0114x + 3.7660 does not tell the trend in preference of Barebones as age goes up. Because the y-axis of preference of Barebones is a finite scale of 1 to 5 (categorical), the regression line equation does not reflect user preference from different ages.

Our conclusion is that from the scatter plot, different ages all rated with varied preferences of Barebones and there is no significant effect of age on the preference of Barebones.