SAN DIEGO STATE
UNIVERSITY

Assignment 3
SAS Visual Analytics

Submitted by

Harshal Sanap
822044776

MIS 691: Decision Support System
Mar 20, 2019

Question 1: What did you do?

- Before starting with the assignments meant for writeups, I followed assignments 1 & 2 from the Teradata network to get acquainted with the SAS visual analytics tool. I watched all their introductory videos and then went to SAS visual analytics cloud platform and selected Insight_Toy_Demo dataset from the data explorer pane for my analysis. I created basic charts in the first two assignments to understand the overall sale of toys in different product lines over different time period. Once I got used to the tool from the first two assignments then I started with assignment 3 & 4.

Assignment 3:

- For this assignment, I started with Report Designer from the SAS VA platform. The assignment required creation of six different graphs from sections like tables, graphs, controls and from others section of the SAS platform. I selected the insight_toy_2017 data set for this analysis.
- **Plot 1**: I was interested in knowing **product sales across different product lines** categorized over their product brands **across different countries**. For this I thought of creating a **pie chart** visual. In report designer I chose the above-mentioned data set and then created one section. In that from objects I selected Pie Chart as my visual and then dragged product line and product sale onto that chart. Later to understand the sales across different product brands I created a **button bar control** by dragging button bar into the control section of the report and then dragging product brands onto that. Finally, to understand how the distribution varies across different countries I repeated the above control step and added a report level drop down list filter for product country. Finally, to make the plot aesthetically appealing, I added the value labels as percentage of total, made them bold and increased the font size from the properties section.
- **Plot 2**: After having checked the product line wise sales distribution, I was interested in understanding the **avg. profit made across different product lines**. For this I first created a **calculated field** by subtracting product material cost from the product sale to get the profit value. Later I set its default aggregation format to average. Now I dragged the bar chart into the section and then dragged the above created profit field and product line onto it. The visual was ready. Later I thought it would be a good idea to check profits across different seasons. For this I created a **custom grouping category** called seasons and added the months corresponding to different seasons in their labels as Fall, Summer and Spring. I then dragged this new category onto the control panel as a button bar.
- **Plot 3**: I was interested to understand the satisfaction of customers related to different product lines along with the total number of sales representatives involved in it. For this I created a **cross tab** and dragged product line on to it as the first index column. Then dragged different measures like sales rep id, unit discard rate, customer satisfaction rate and made their default aggregation type as average except for sales rep id it was made count to get the total count of sales rep involved per product line.
- **Plot 4**: I wanted to setup **gauges** to understand if the **sales target is met or not** across different product brands. I set up a target of $3Million across both the brands. For this I dragged the gauge from other section of objects onto the report. Later I dragged product brand and product sale onto the gauge. Finally, I **created a display rule** and setup 3 intervals with target amount as $3Million.
- **Plot 5**: I was interested in understanding the cities involved behind most frequently placed orders and wanted to mark the avg. product price for that city. For this I created a **word cloud** by dragging the word cloud visual into the section and then dragged city onto it. The size of words was set to word frequency and I also added color of words depending upon average price per city from properties section of the visual.
- **Plot 6**: To understand if the **unit discard rate improved** over time period or not, I created a line chart. For this I dragged line chart into the section and then dragged unit discard rate and month

onto it. Later to understand how the rate is different over product brands and product lines I created a **hierarchy of product brand** and product lines under it. I then dragged this hierarchy over the group section of the visual. This allowed to **drill down** the product brand wise discard rate to product line wise rate. I finally dragged **text** title onto the report to name the report as "Overall Unit Discard Rate Change Over Months".

- Finally, to the above report I added a **global level filter** of product brand. This would allow to **slice** the whole section output to product brand selected in the filter. It was done simply by dragging a button list object onto the report level filter section and dragged product brand onto it.

Assignment 4:

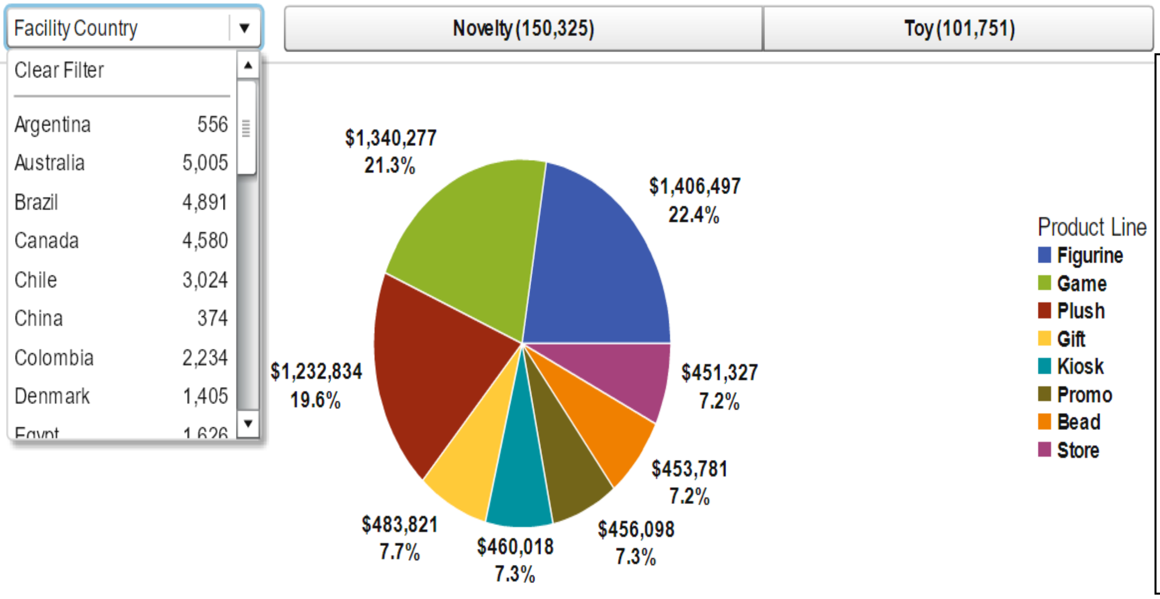
I used data explorer for the hypothesis testing and the report designer for the task after that.

- **H1:** To understand how closely the weight levels and cholesterol are related, I first researched on how a categorical variable can be compared with continuous variable. I realized that we can write a statistical code in SAS and use multiple logistic regression, but I couldn't find a place to write the code on SAS cloud platform. Thus, I came up with a new approach where I plotted a bar chart of weight level against their **avg. cholesterol** measures. As there was a missing entry in weight level, I created a filter and removed the single missing category from the bar chart.
- **H2:** To understand if men are usually more obese than women, I first thought of plotting the obesity rate or obesity measure for both the sexes but as that rate was not given or a formula for the same was unknown thus I checked on google on how to understand if a person is obese or not given their weight and height. That lead me to **BMI** calculation, I created a **calculated field** called BMI as (weights in pound * 0.45)/(heights in inches * 0.025) ^2. Later I created a **new binary categorical variable** that will output 1 for BMI >30 meaning obesity and 0 for BMI <30. Finally, I plotted this new category against the gender to understand the overall frequency of obese individual in both the genders.
- **H3:** This hypothesis stated that Women usually smoke less than men, but their cholesterol level is higher. For this I first created a **new category** as smoker or not which had two labels first for all the smokers taken from smoking status as light, heavy, very heavy or moderate and second label as non-smoker who didn't smoke at all. The I created a **bar chart** of gender along with the newly created category and I plotted the **percentage distribution** of the people who smoke in that population. This gave us the results to understand the first part of the hypothesis. For the second part I created another visual and plotted average cholesterol level across both the genders.
- **H4:** Hypothesis stated that the blood pressure is higher for people with higher cholesterol level. To test this, I simply plotted different blood pressure levels against average value of their cholesterol in the form of a bar chart.
- To make some insights related to the population suffering from coronary heart disease, I made a **global filter** by dragging cause of death as drop-down list global filter. Then I selected CHD from it. Now I wanted to understand how different variables are impacted for a population with CHD, for this I first plotted a bar chart for gender to understand the gender-wise frequency distribution in CHD population. Then I created pie charts for weight and blood pressure status to understand how statuses impact CHD population. Later I checked the impact of smoking on CHD by plotting a bar chart of smoking status and getting the count of people with CHD in each smoking status. Finally, I created a **dual axis** chart to understand the average age when CHD is diagnosed and average age of death for both the genders. I created the dual axis chart by dragging dual axis bar chart on to the section followed by gender as axis and age at death and age at CHD diagnosed as measures. Later I made the default **aggregation** for the measures as average.

Question 2 : What were the results?

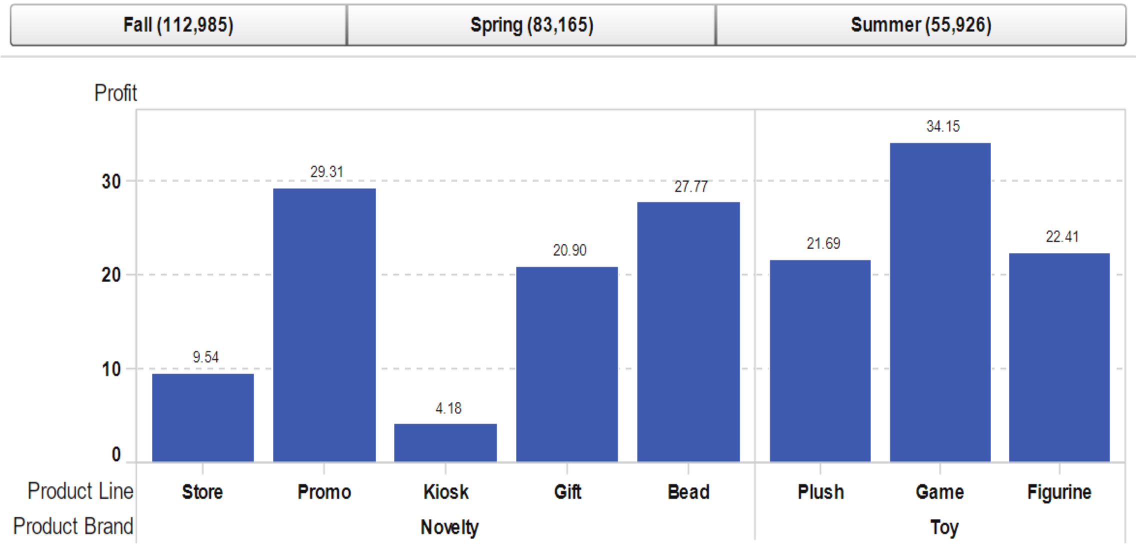
Assignment 3:

- Plot1 : The product line of **Figurine** is contributing to the **highest product sales** amount of almost $1.4 Millions **followed** by **Game** product line. From the below figure we can also see that there are **higher number of Novelty related products** being sold than toys but toys are contributing higher towards total sale, which means that they seem costlier than novelty items. Toy product line includes Game, Figurine and Plush. The **button bar control** section provides **dicing** feature to see the overall sales specific to product brand while the **country drop down** provides **slicing** feature on the report to filter the results for a specific country.



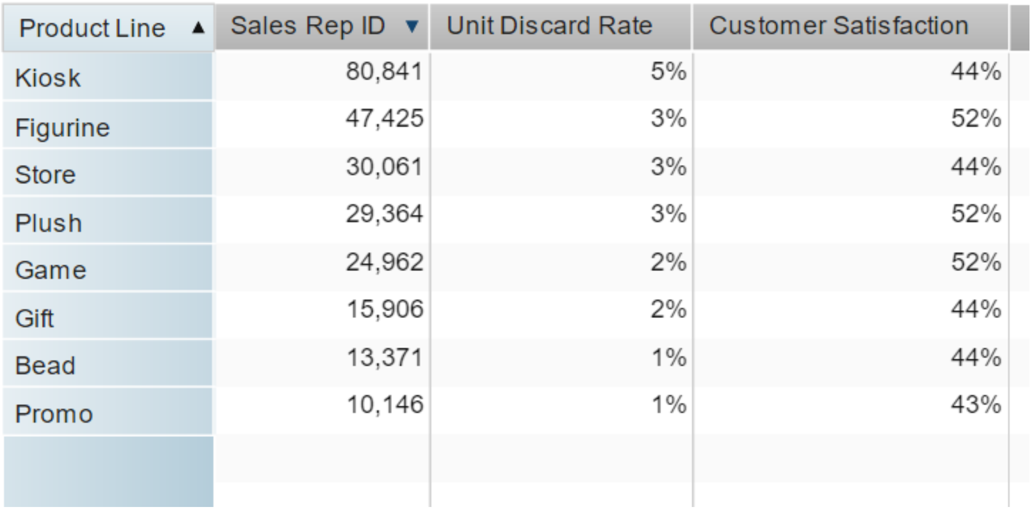
Visual 1: The Visual highlights the product sale distribution across different product lines in the form of a pie chart. The labels indicate the total amount of product line sold in $ and the total percentage contribution overall. The drill down can be used to understand this distribution across different countries. And the buttons to specify sales across particular product brand.

- Plot 2: **Average profits** are **highest** for product line **game** as part of the brand toys **followed** by **promo** of line Novelty. The highest number of products are also sold during Fall season followed by spring and surprisingly not summer.



Visual 2: The Visual is a bar chart with x axis showing product lines grouped under the product brands. And y axis shows the average profit across those product lines. The buttons on top highlight the different sale seasons which can be used for slicing the data to that specific season.

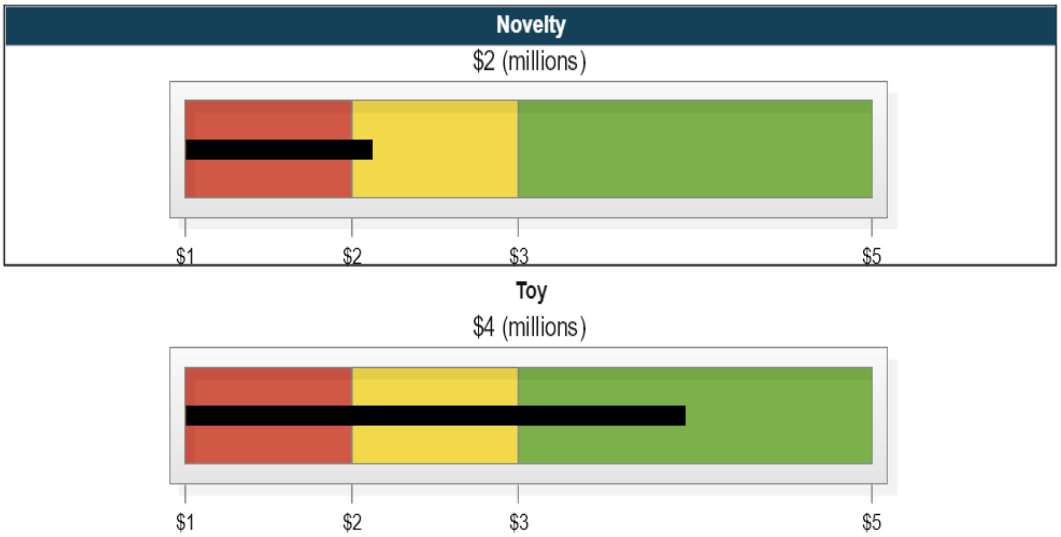
- Plot 3: From the below table we can understand that **kiosk** has the **highest discard rate** on an average with a **very low customer satisfaction**. This explains the reason behind making it being hard to sell and having so many sales rep involved in it.



Product Line ▲	Sales Rep ID ▼	Unit Discard Rate	Customer Satisfaction
Kiosk	80,841	5%	44%
Figurine	47,425	3%	52%
Store	30,061	3%	44%
Plush	29,364	3%	52%
Game	24,962	2%	52%
Gift	15,906	2%	44%
Bead	13,371	1%	44%
Promo	10,146	1%	43%

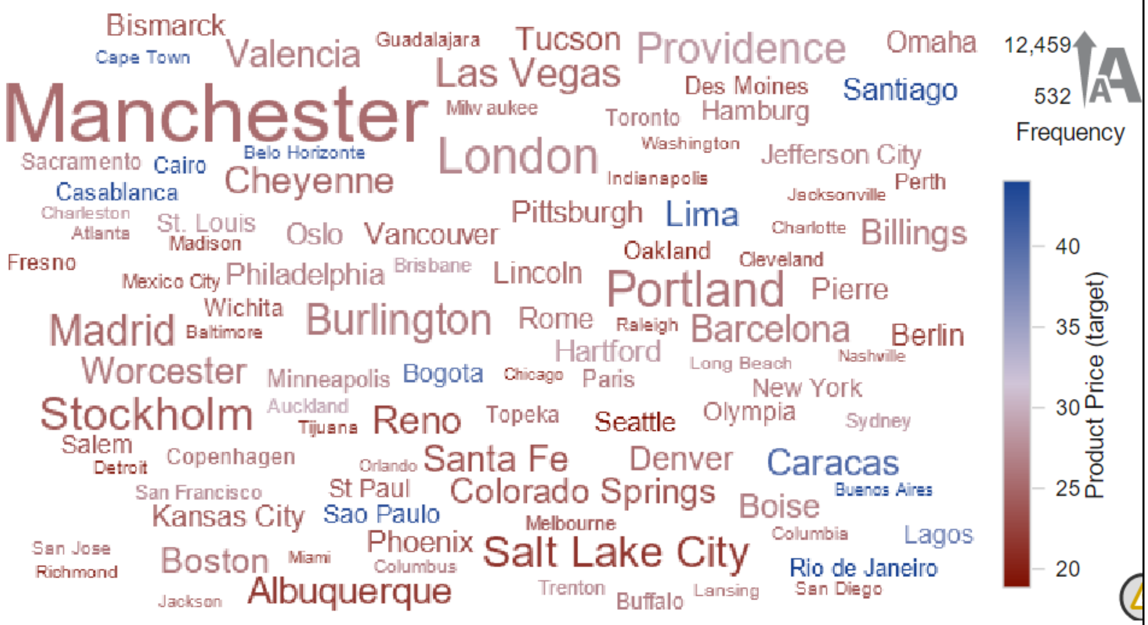
Visual 3: The cross-tab visual has product line as the index column. And for individual product lines we have the count of sales rep, avg unit discard rate and avg customer satisfaction as the measures.

- Plot 4: From the below plot we can see that the **toy product brand** has **reached** the product **sales target** of $3 millions and Novelty brand is still left with one million to reach the target.



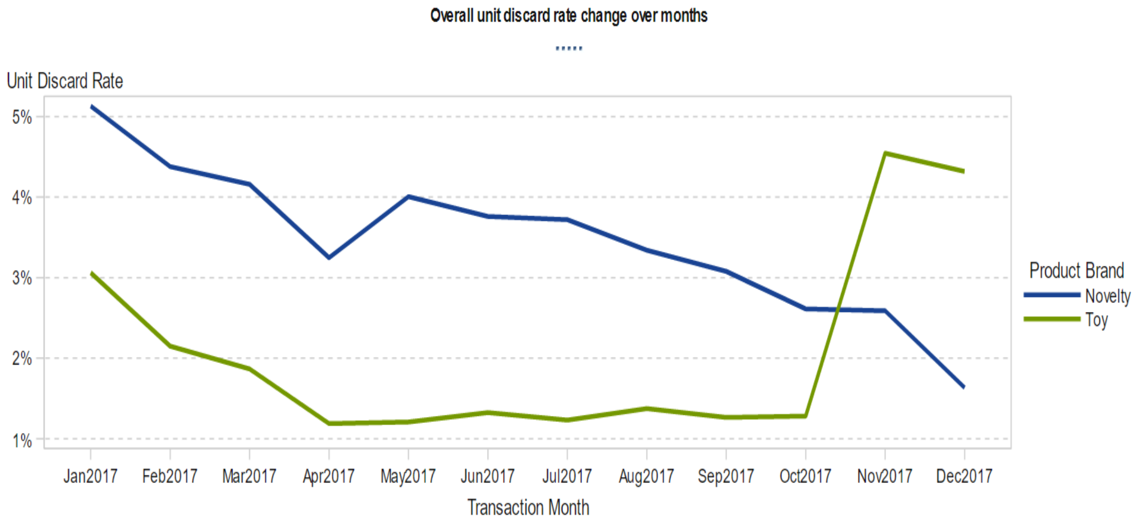
Visual 4: The gauge visual helps us understand how far we from the target are. The beginning of green part represents the target which is $3 million and yellow part means we are close to our target and above $ 2 million. Red meaning, we are far from target by at most $2 million. For Novelty we can see the black bar is in yellow meaning it is close to target by at most $1 million and for Toy it is in green meaning it has reached the target of $3 million.

- Plot 5: From the below **word cloud** we can understand that **the highest number of orders** are being placed from **Manchester** and that's why its size is shown the biggest. However, Buenos Aires has the highest avg price per product and that signifies the order number frequency with that city being low and that's why its size is least in the below cloud.



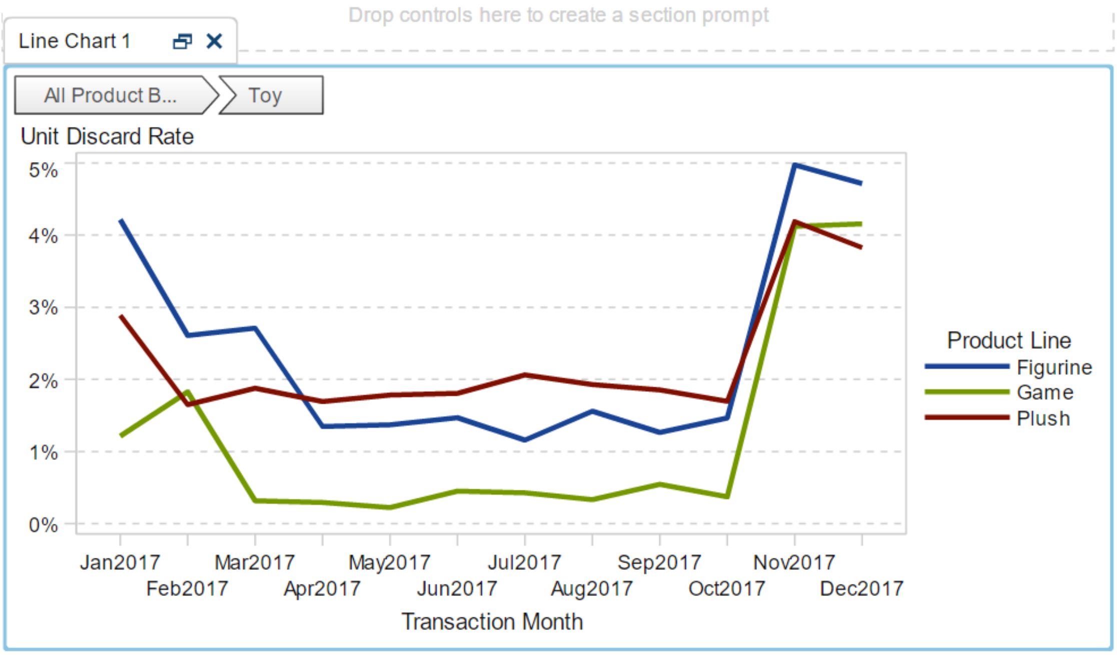
Visual 5: The Word Cloud has city names as the words present in it. Its size is dependent upon the number of orders being placed from that particular city. If large number of orders are being placed from a city then it's size will be big compared to the ones with lower order frequency. The color is dependent upon the avg product price. The legend is displayed on the right. The frequency legend tells us that the lowest frequency is 532 and highest being 12,459 and least avg. product price is around 20 which is in red color and highest avg. price is around 40+ which is in blue color.

- Plot 6: From the below plot we can see that for **Novelty product** there is a **declining trend** from start of the year till end which looks good as it signifies that lower products are being discarded as the time passes and it shows improvements in the product or service. For Toy product line the discard rate dropped till October and because of some reasons it jumped up in November and thus we need to investigate it further on it.



Visual 6: The line chart displays the trend of change in unit discard rate over month. The x-axis represents months and y axis represents the unit discard rate %. On right there is a legend of product brand with Novelty in blue and Toy in green.

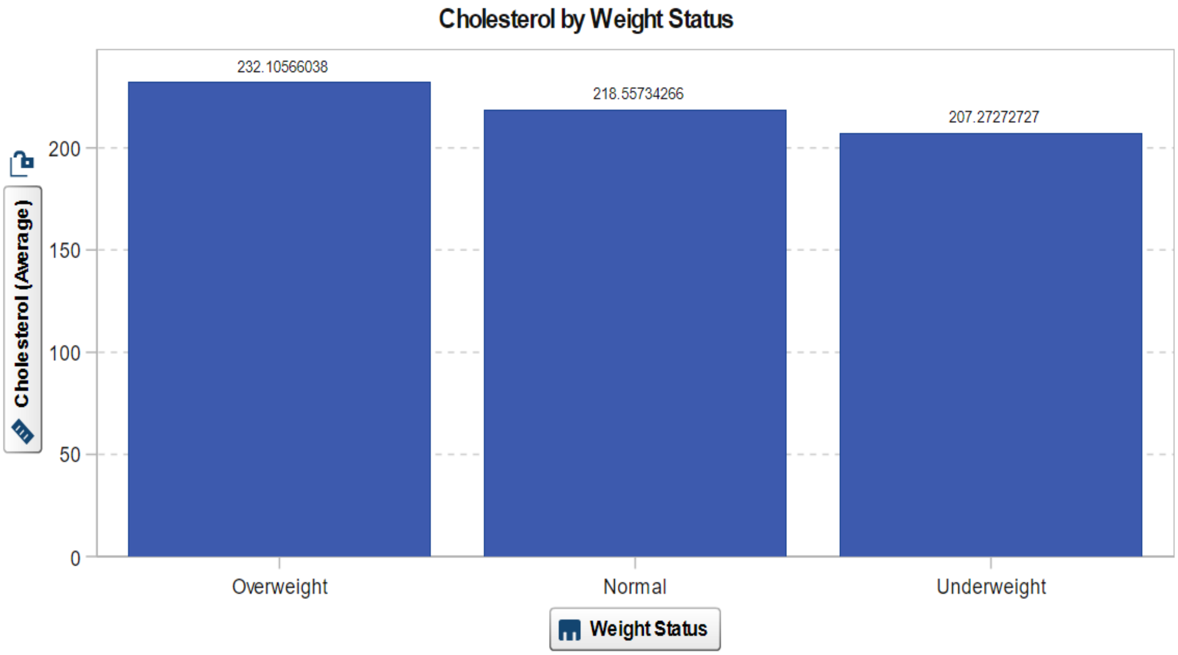
- The above visual is **drilled down** to consecutive product lines after clicking on the product brands. From above chart I was curious in understanding where and why exactly there was increase in the Toy product brand. Thus, I drilled down on it and from the below chart I understand that there was an increase in all product lines present in Toy product brand.



Visual 7: This is the drilled down version of the above visual. Product brand toy is drilled down to its 3 individual product lines along with their individual discard rates.

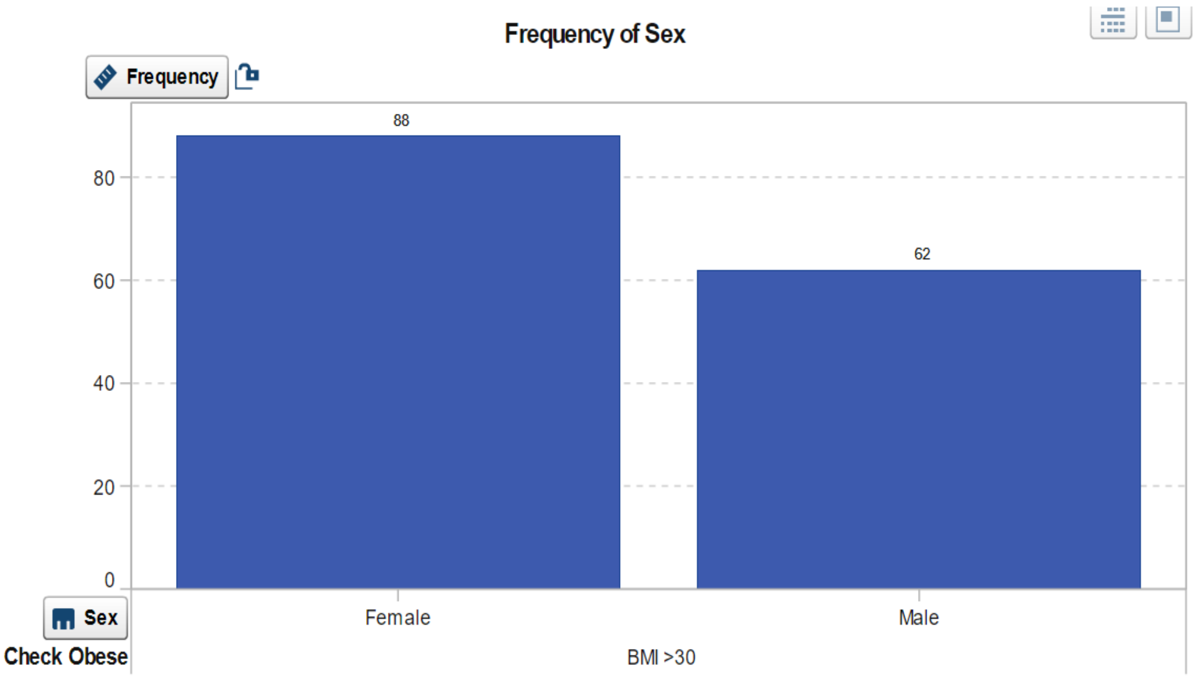
Assignment 4:

- H1: The average cholesterol level is the highest for overweight status followed by normal weight status. Thus, our **hypothesis** is **true,** and we can say that there does exist a some relation between weight status and the cholesterol level for different measures.



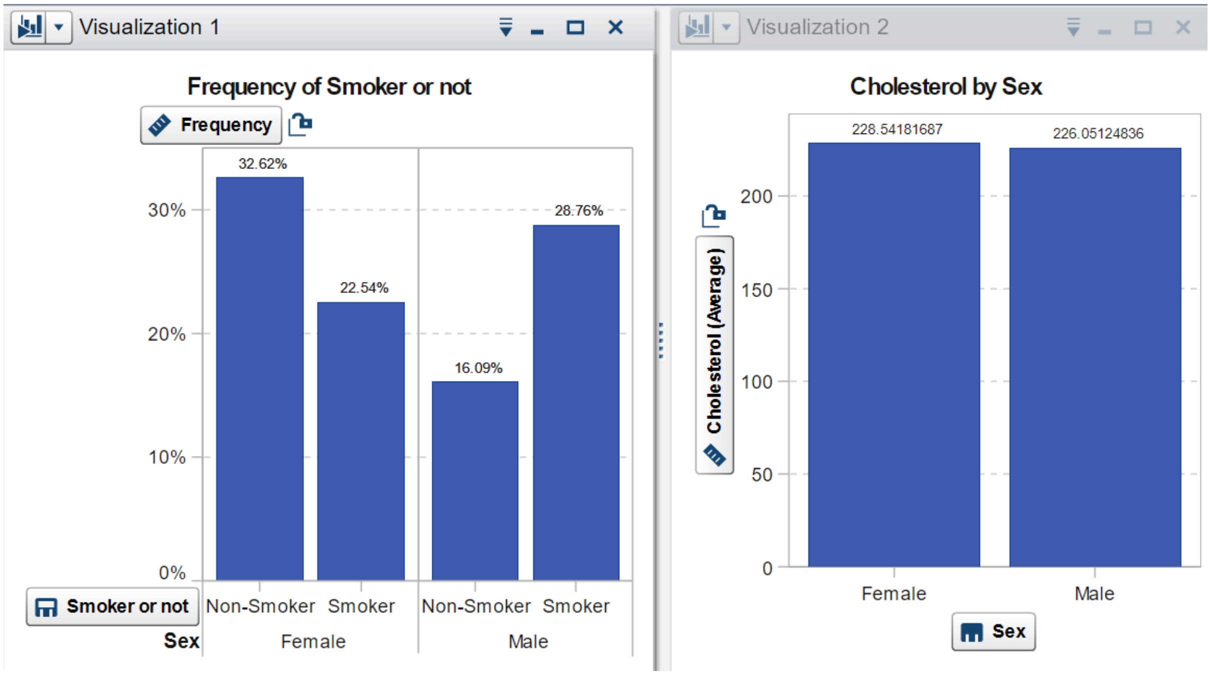
Visual 8: This bar chart has different weight status on x-axis and average cholesterol levels on y-axis.

- H2: Our **hypothesis** of men tends to be more obese(BMI>30) than women is **false** as from below chart we can see higher number of women are obese compared to men.



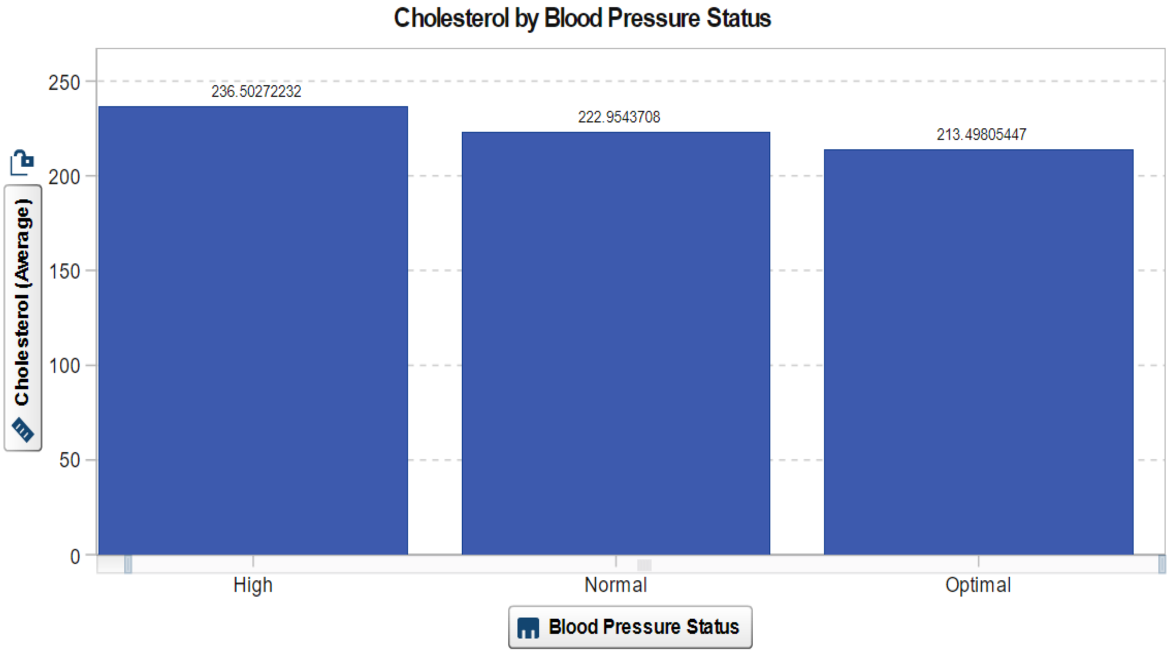
Visual 9: This bar chart has the frequency of total number of obese individuals in both the genders. On x axis we have gender and y axis has the frequency of individuals being obese.

- H3: The **hypothesis is true** that female tend to smoke lesser than men but still their cholesterol content is higher. There is only 22.5% of the population from females that smoke compared to almost 29% smokers from males. The cholesterol level on average in females is 228.5 compared to 226 for males.



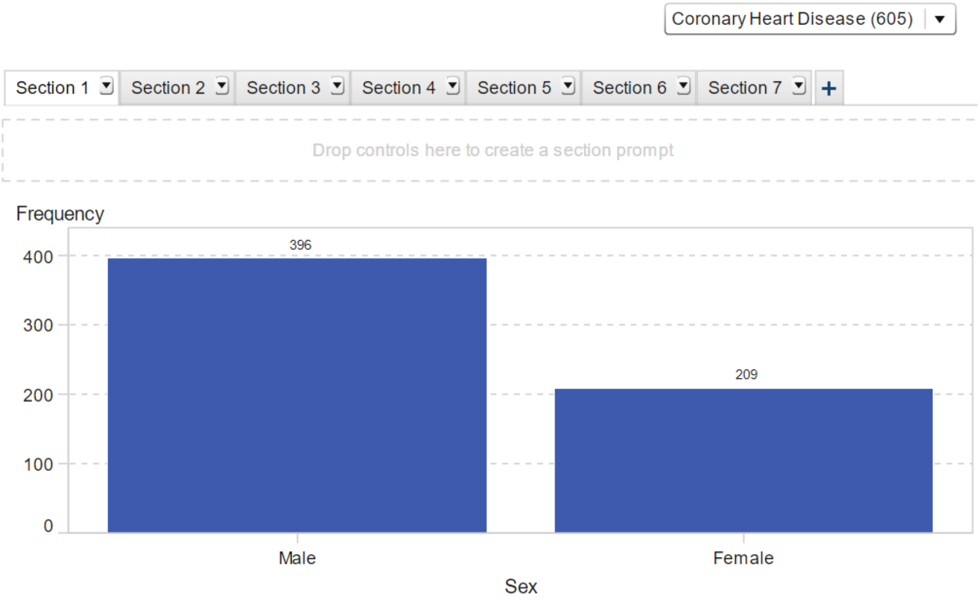
Visual 10: The first visual tells us the percentage of population who smokes in both the genders. The visual on the right tells us the average value of cholesterol in different genders.

- H4: The **hypothesis** of the blood pressure is higher for people with higher cholesterol level is **true**. As seen in the below chart individuals with higher blood pressure tend to have higher on average cholesterol level thus it proves the hypothesis true.



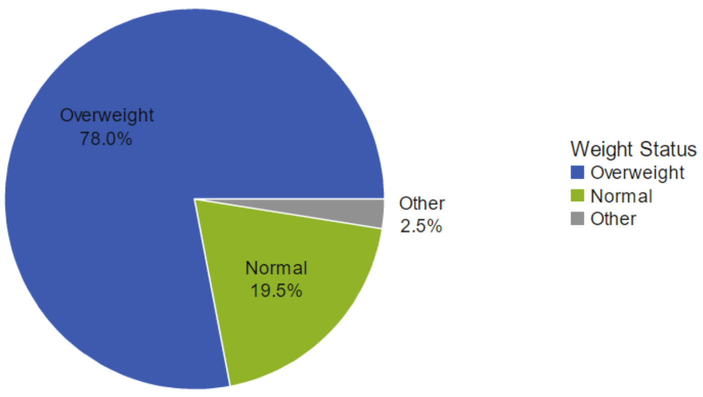
Visual 11: The visual is a simple bar plot with levels of blood pressure on x-axis and average cholesterol levels on y-axis.

- Characteristics of people who suffered from CHD can be discussed as below. Out of the total population, 605 people suffered from CHD. Males tend to be more prone to CHD than females.



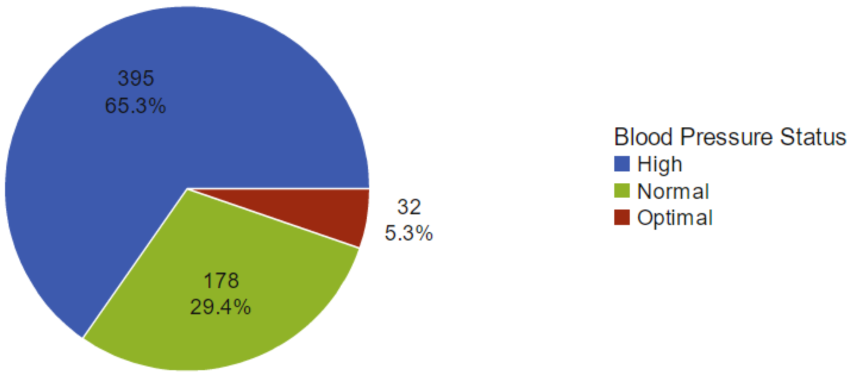
Visual 12: The visual is a simple bar plot sex on x-axis and number of people in that gender having CHD is given on y-axis . The global filter is set as population with CHD.

- **Weight has a very big impact** on whether you will **get CHD or not**. Almost **80%** of the population who were overweight suffered from CHD.



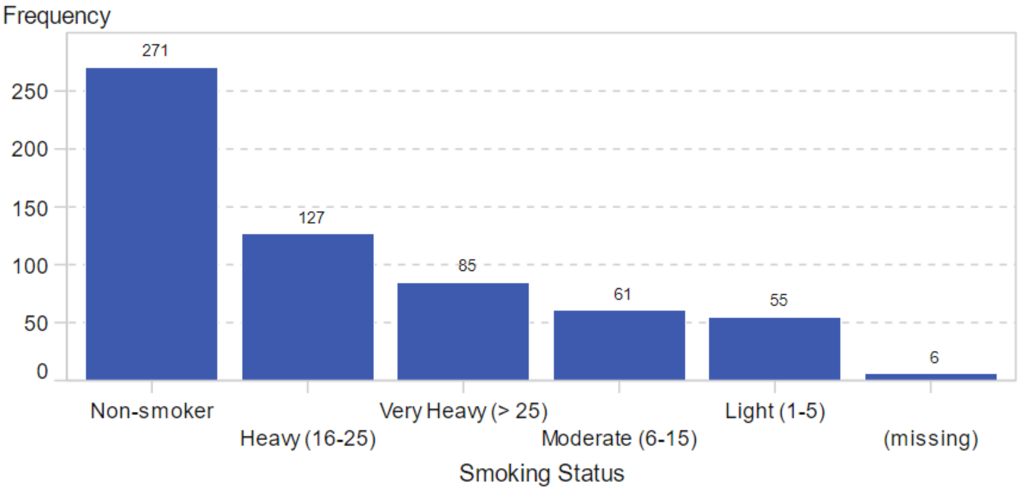
Visual 13: The visual is a pie chart showing percentage of population suffering from CHD at different levels of weight.

- **High blood pressure** can also **contribute to CHD**. Almost 65% of the people who had high blood pressure suffered from CHD.



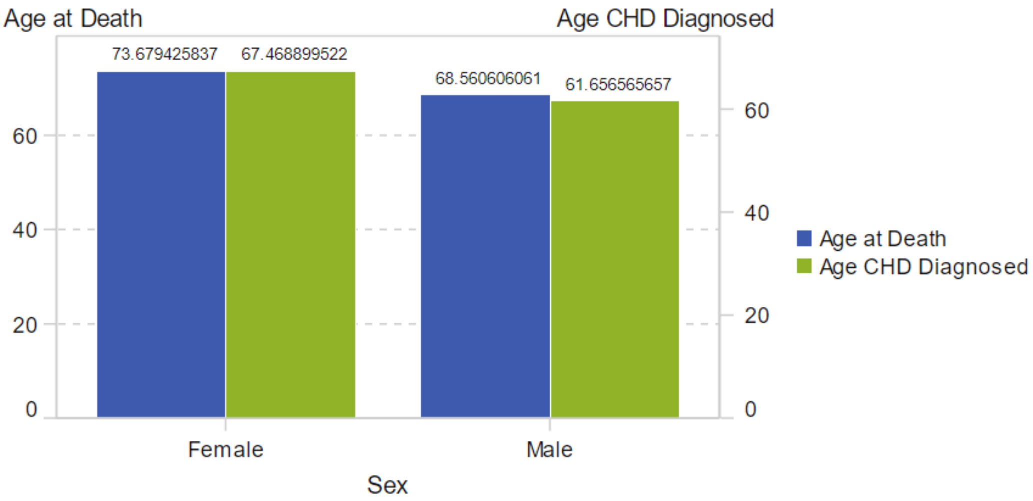
Visual 14: The visual is a pie chart showing percentage of population suffering from CHD at different levels of blood pressure.

- Surprisingly, **smoking doesn't seem to be contributing that significantly towards CHD**. As almost 45% of the people who didn't smoke had CHD.



Visual 15: The visual is a bar chart showing frequency distribution of CHD population across different smoking statuses.

- The average age of females when they get diagnosed with CHD is almost 6 years higher than that of males, but male tend to survive almost an year longer than females after being diagnosed with CHD.



Visual 16: The visual is a dual axis bar chart with sex on x-axis and two columned measure as avg age CHD diagnosed and avg age of death on y-axis.

Question 3 : What did you learn?

- I cleared my concepts related to **slicing** and **dicing** after implementing it again in **plot 1** of assignment 3 and I also learnt more about **hierarchy** along with **drill down** operations with **plot 6** of the same assignment.
- I personally found SAS to be extremely slow and not so intuitive compared to Tableau or MicroStrategy or PowerBI. And I would not prefer SAS over others for **visualization**.
- I didn't like the idea of dragging the objects related to visualization type on the section first for creating a visual in SAS report designer. I believe just dragging the categories or measures should have prompted it directly and then if required then we could have updated the visual manually.
- I **liked** different **Business Intelligence operations** like Year Over Year growth, year to date growth etc. available directly from the measures options. Compared to Tableau or PowerBI where you must write a small code to perform some BI operations.
- I also personally feel that doing SAS VA before Tableau would have been better as tableau looks like a big upgrade after SAS.
- SAS **doesn't allow undo** option which can also be a problem at times, this is however not the case with PowerBI or Tableau.
- SAS has data explorer which is meant for quick exploration of data, more of like pivoting analysis of excel and they also have reporting environment which can be used for quick hypothesis testing. It also has different statistical models like logistic/linear regression, GLM, decision tree etc.
- I didn't like separate sections for data visualization and data. I prefer them to be together the way it is given for Tableau or PowerBI.
- Deleting visuals or contents from the visual required multiple steps in SAS which is time consuming.
- From my previous exercises I had a **good understanding of creating visualizations** and thus it was quite intuitive to make few charts on my own.
- Image was not added onto the report as it was taking too much space out of the report visual reducing the size of the visual.
- Though I have highlighted some of the features that I didn't like about SAS, I liked the statistical manipulative powers of the tool and its ability of forecasting and running correlations or logistic regressions.
- From this exercise I learnt about testing a hypothesis through visuals rather than running some statistical test in R.
- From this exercise I also learnt that data can sometimes tell a different story that what we imagine.

Question 4: How does it relate to class?

- From Night 6 from presentation **jennex_intelligence_strategy**, **slide 3** talks about the **actionable insights** and **slide 4** provides the organizational learning pyramid. These two slides helped me understand how the process of information gathering and decision-making works. I can also relate the steps mentioned in **slides 5-9** to my work as follows: Step 1: Define the decisions, it is like our definition of hypothesis, step 2: identify the key knowledge needed to support actionable insights, its like understanding the basic heuristics to support the hypothesis for example obesity is related to BMI and we have a formula to calculate BMI and BMI >30 is called obese, step 3: identify the needed information, its like identifying the data to support our hypothesis in our case we used Heart, step 4: identify tools and filter to obtain data and information, its like identifying tools to visualize and manipulate the data, in our case that would be SAS, finally the last step is to make visualization and test the hypothesis to make support it in decision making.

- The presentation of **Marakas_Ch08_Knowledge_Engineering** from night 6 talked about differentiating data, knowledge and information and how knowledge is acquired. I was able to relate this to this assignment where I gathered and analyzed data for knowledge discovery.
- On night 4, the **slides 2-5 and 24** of **sharda_dss10_ppt_04** taught me about business reporting and how to make a successful report. From this I thought of adding data labels and formatting in visuals created above. **Slide 13** from the same ppt also taught me different types of graphs and charts that are possible.
- **Slides 22-33** from **DSS_User_Interface_Dialogue_and_Visualization_Design** helped me understand how to make a right visual for the given audience.
- I also referred to the **book Business Intelligence and analytics**, chapter 4, section 4.2 helped me understand the importance of business reports and different types of report, section 4.4 taught about different types of graphs and how to apply them to the dataset, section 4.5 talked about SAS platform.

References:

- https://www.teradatauniversitynetwork.com
- https://www.gartner.com/reviews/market/analytics-business-intelligence-platforms/compare/sas-vs-tableau