**Name:** Rajat Sethi

**Objectives:** The purpose of this project is to introduce the statistical software JMP. Upon successful completion of this project, you will be able to…

* Create graphical displays and calculate statistics in JMP
* Describe a distribution
* Use the output provided by JMP to answer relevant questions
* Select a random sample
* Understand the possible effects when a random sample is not used in a study
* Use JMP to draw two box plots on the same scale.
* Compare two distributions
* Compute probabilities using a table of values

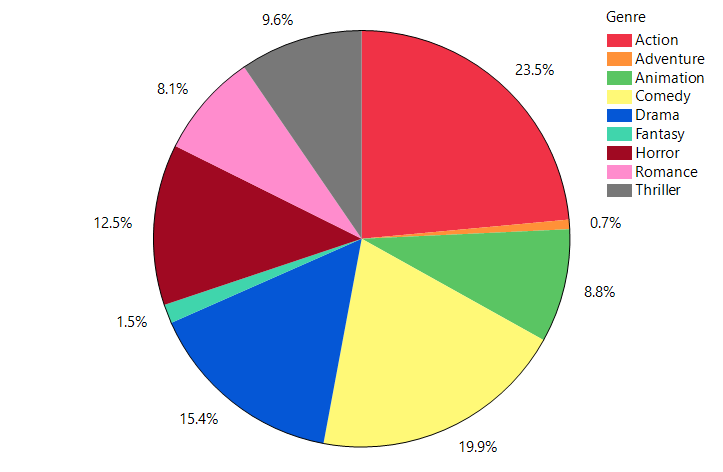
Delectable Delights is a large consumer food manufacturer selling its products in retail stores nationwide. You have landed your first job after graduation from Clemson in their advertising division. Since you took statistics as a part of your coursework, you are often called upon to perform data analysis for the advertising division, as well as other divisions of the company.

**Directions:** Answer the following questions using complete sentences as though you were presenting your analysis to the employees of Delectable Delights. Please provide any appropriate output and/or screenshots from JMP. Instructions for creating several types of graphs or tables and statistics can be found on Canvas in the file **JMP Instructions.docx**. Paste your answers and any output into this document. This project is worth 100 points.

1. (20 points) Delectable Delights would like to feature some of its products in major motion pictures. To assist in selecting possible move projects to approach, your boss, Ray Holtz, would like you to review movie genres to see which is most popular. Since Delectable Delights uses the software JMP, you decide to look into the Hollywood Movies data set that comes with JMP. You can find the file **Hollywood Movies.jmp** on Canvas.

Using the column “Genre” in the Hollywood Movies data set, create a graphical display in JMP that highlights the proportions of the different genres represented and report to Ray on the one that is most popular. The type of graph you choose needs to quickly convey the proportion or percentage of each genre. (Hint: Consider the type of data and select a simple, powerful way to illustrate the proportions of genres represented. There is more than one type of graph you could use.

* 1. Copy and paste your graph here. Make sure to include titles or legends. (10 points)
     1. If you are using the **Graph Builder** function in JMP, you can change some of the titles by double-clicking on the name given by JMP and typing your own title. You can make changes to the legend by double-clicking in the region where the legend is located.
     2. If you have selected **Analyze >> Distribution** to create your graph, you can double-click any of the titles to change the names. For example, you can change the title of the column **Prob** to **Proportion** or **Relative Frequency** since that column represents the relative frequency of each genre. Double-clicking on the horizontal axis allows you to make a variety of changes to that axis. Clicking the red triangle next to the name of the variable gives you other options, including the ability to order your data from largest to smallest categories (or vice versa).
     3. Explore the various options available in JMP to change the look of your graph or table. There are also many videos available on YouTube on how to use JMP.



* 1. Explain your choice of graph for Part (a). (5 points)

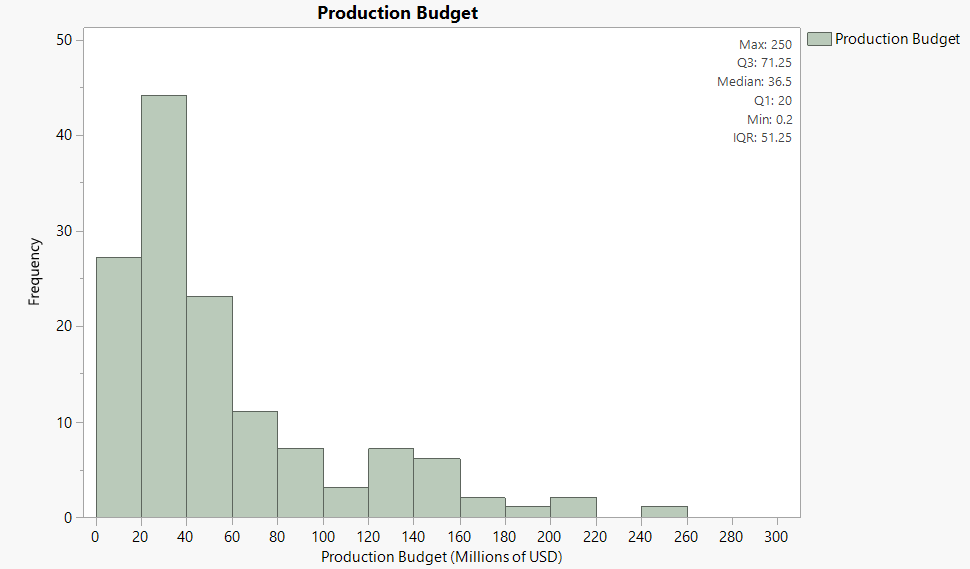
I used the Pie Chart because it fits the requirements. It quickly shows the percentage of each genre from the sample data. Larger slices represent larger proportions.

* 1. Using your graph or table from Part (a), report to Ray the most popular genre. Explain your answer based on the graph or table. (5 points)

Action movies are the most popular, as they have the largest proportion taking up 23.5% of the sample data.

1. (15 points) Joan Shilling, who is Ray’s supervisor, would like a graph of the movie production budgets. Ray asks you to consider the Hollywood Movie data set and to create a histogram using JMP that highlights the shape of the distribution of the production budgets and to provide a measure of center and spread.  
   1. Use JMP to make a histogram and to calculate a measure of center and spread of the movie production budgets in the Hollywood Movie data set. The units for the production budgets are in millions of dollars. Include labels, a title, units and other useful information in your graph. The measure of center and spread you select should be based on the shape of the distribution. Paste your histogram and statistics here.

You may investigate display options and the bin width options and pick ones that you find visually pleasing when creating the histogram. You should also add a horizontal axis label. To add the label, right-click on the horizontal axis and choose “Add Axis Label…” There are YouTube videos on how to change he appearance of histograms in JMP that you can use to help you make your choices. (10 points)



* 1. Using the information from Part (a), describe the distribution of the production budgets for Joan. (5 points)

The distribution is skewed to the right.

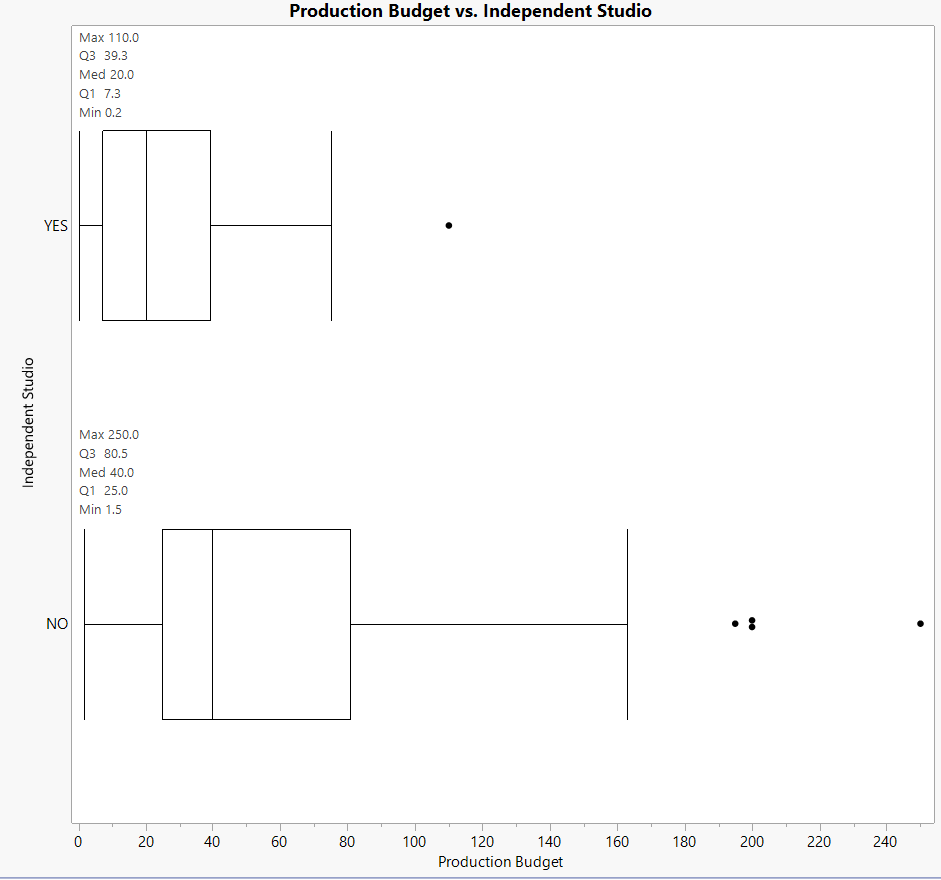
The median production budget is $36.5M.

The middle 50% of movies range between $20M and $71.25M.

The smallest budget in the sample data is $0.2M.

The largest budget in the sample data is $250M.

1. (15 points) Frank Holland, who works with Joan Shilling, wonders if there is a difference in the production budgets for films produced by major studios and those produced by independent studios. You are asked to make two box plots for the production budgets, one for independent studios and one for major studios, and to compare the two budget distributions.
   1. Make two box plots in JMP’s Graph Builder. Go to **Graph >> Graph Builder**. Drag the column titled **Production Budget** to the x-axis and drag the column titled **Independent Studio** to the y-axis. Select the graph symbol at the top that resembles box plots. You will now have two box plots that you can easily compare. On the left select **5 number summary**, and JMP will add the statistics to your graph. (Recall that JMP will calculate the quartiles differently than we did in class. Simply use the JMP output, you do not need to calculate the quartiles by hand.) Place the graph of the box plots in the space below. Add a title in if JMP doesn’t produce one for you. (10 points)



* 1. Write a short paragraph to Frank comparing the two distributions. Make sure you compare the shape, center, spread, and unusual values if appropriate. (5 points)

Non-Independent studios have consistently larger production budgets than their counterparts.

They have a larger:

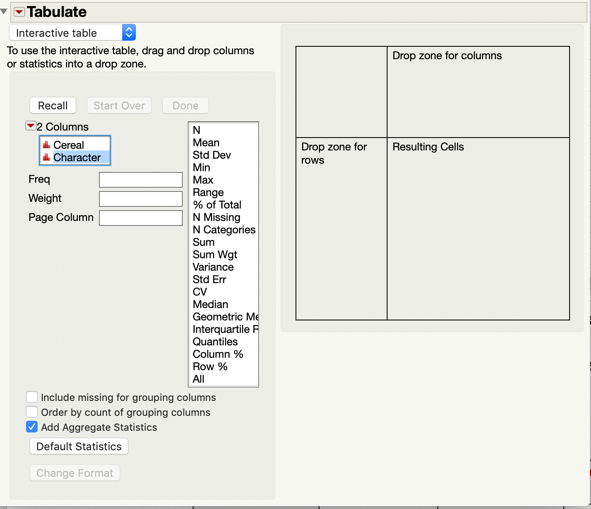
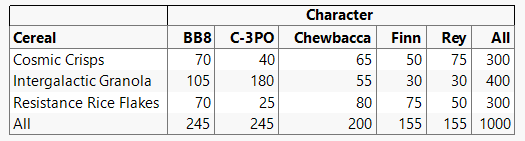
* Minimum
* Q1
* Median
* Q3
* Maximum
* Range
* IQR

They also have more outliers than the independent studios.

Both distributions are skewed to the right.

1. (25 points) Delectable Delights is very excited to announce that they will provide advertising support for the Star Wars Episode IX movie. They will feature 5 of the Resistance characters as toy prizes in their cereals *Cosmic Crisps*, *Intergalactic Granola,* and *Resistance Rice Flakes*.

A shipment of 1000 cereal boxes has been prepared for distribution. The file **Star Wars Promotion.jmp** contains the data about the cereal names and which toys are included in the boxes.

1. Using the Tabulate function in JMP, create a table showing the 3 cereals and the 5 types of characters that are in the boxes. Paste the table below. Instructions on how to use the tabulate feature can be found in the **JMP\_Technology\_Guide** file on Canvas. However, since there are two variables (Cereal and Character), drag and drop one of the variables into the section labeled **“Drop zone for columns”** and the other into the section labeled **“Drop zone for rows”.** Also select **“Add Aggregate Statistics”** so that you get column totals. Once JMP has made the table, you can change any of the row or column headings if you wish. Place a copy of your table below. (5 points)
2. Suppose you randomly select one box of cereal from this shipment. What is the probability that the prize is the character Chewbacca? Use probability notation when you write your answer. You may write your final answer as a reduced fraction or as a decimal. (5 points)

P(Chewbacca) = 200/1000 = 1/5 = 0.2

1. Given the event that you selected a box of *Resistance Rice Flakes* from the shipment, what is the probability that your prize is the character Chewbacca? Use probability notation when you write your answer. You may write your final answer as a reduced fraction or as a decimal. (5 points)

P(Chewbacca | Resistance) = P(Chewbacca AND Resistance) / P(Resistance) = 80/300 = 4/15 = 0.266666…

1. Are the events selecting a box of *Resistance Rice Flakes* and getting the character Chewbacca as the prize independent? Justify your answer using probabilities. (Hint: Use a test for independence from the Chapter 5 notes in the lecture guide.) (5 points).

P(Chewbacca | Resistance) = 0.26666…

P(Chewbacca) = 0.2

Since P(Chewbacca | Resistance) =/= P(Chewbacca), they are NOT independent.

1. Are the events selecting a box of *Resistance Rice Flakes* and selecting the character Chewbacca mutually exclusive? Justify your answer using probabilities. (Hint: Use your Chapter 5 notes.) (5 points)

P(Chewbacca AND Resistance) = 80/1000 = 0.08

Since P(Chewbacca AND Resistance) =/= 0, they are NOT mutually exclusive.

1. (25 points) In this question, we will explore a possible effect of using a self-selected sample (which is a type of convenience sample) rather than a random sample. Take a look at the list of the 200 top grossing movies of 2018. The list can be found on Canvas in the file **Top 200 Movies of 2018.docx**.
   1. Select a sample of 10 movies that you saw (or wanted to see) in theaters and write the titles of those 10 movies in the table below, along with the amount they grossed in 2018. This is your self-selected sample. Notice that the listing of movies gives the gross income rounded to the nearest million, so that a movie listed as earning $191.5 really grossed $191,500,000. The order in which you write the movies in the table below does not matter. (5 points)

|  |  |
| --- | --- |
| **Movie Title** | **Gross Income (Millions)** |
| Avengers: Infinity War | 678.8 |
| Deadpool 2 | 318.5 |
| Venom (2018) | 213.5 |
| Ralph Breaks the Internet | 201.1 |
| My Hero Academia: Two Heroes | 5.8 |
| Ready Player One | 137.7 |
| Ant-Man and the Wasp | 216.6 |
| Black Panther | 700.1 |
| Spider-Man: Into the Spider-Verse | 190.2 |
| The Mule | 103.8 |

* 1. Compute the average gross income for the sample of movies you selected. Show how you computed the average and include units in your answer. (5 points)

Average Gross Income = (678.8 + 318.5 + 213.5 + 201.1 + 5.8 + 137.7 + 216.6 + 700.1 + 190.2 + 103.8) / 10 = $276.61M

* 1. Now you will select a random sample of 10 movies. Open **Microsoft Excel** and enter the formula **=RANDBETWEEN(1,200)** into 10 cells. This will generate 10 random integers between 1 and 200.

Alternatively, you can go to the website Random.org to generate 10 random integers between 1 and 200. The link to the random number generator is: <https://www.random.org/integers>. Record the random numbers in the ID column in the table below. If you have a repeated number, generate another random integer so that there are no duplicates in the ID column. Select the 10 movies from the population of 200 movies that correspond to these ID numbers. Record the Movie Title and Gross Income of these 10 movies in the table below. (5 points)

|  |  |  |
| --- | --- | --- |
| **ID** | **Movie Title** | **Gross Income (Millions)** |
| 74 | Love, Simon | 40.8 |
| 144 | A.X.L. | 6.5 |
| 73 | Paddington 2 | 40.9 |
| 5 | Aquaman | 335.1 |
| 100 | Overlord | 21.7 |
| 80 | Peppermint | 35.4 |
| 96 | On the Basis of Sex | 24.6 |
| 31 | The Equalizer 2 | 102.1 |
| 30 | The Mule | 103.8 |
| 132 | Traffik | 9.2 |

* 1. Compute the average gross income for the random sample of movies. Show how you computed the average and include units in your answer. (5 points)

Average Gross Income = (40.8 + 6.5 + 40.9 + 335.1 + 21.7 + 35.4 + 24.6 + 102.1 + 103.8 + 9.2) / 10 = $72.01M

* 1. We often use statistics calculated from samples to estimate the true parameter value of a population. In this case we are considering the average gross income for the population of 200 movies, which is $55.1 million.
     1. Which of your samples came closest to the true average? (2 points)

The Random Sample

* + 1. Regardless of your answer in part i, when using the self-selected sample, are you likely to over-estimate or under-estimate the average gross? Why?

(3 points)

I was likely to over-estimate because I only watched the popular films with higher ratings. Those hyped-up, high-quality, heavy-advertised films likely had larger budgets to garner all that attention and success.