Assignment 1

Problem-1: Using ‘VehicleFailureData’ available on the course website, answer the following:

a. Classify each variable as nominal, ordinal, interval or ratio.

1. `Vehicle Number` - Nominal
2. `Failure Month` - Ordinal
3. `Mileage at Failure` - Ratio
4. `Labor Hours` - Ratio
5. `Labor Cost` - Ratio
6. `Material Cost` - Ratio
7. `State` - Nominal

b. Identify qualitative and quantitative variables.

1. `Vehicle Number` - Qualitative variables
2. `Failure Month` - Qualitative variables
3. `Mileage at Failure` - Quantitative variables
4. `Labor Hours` - Quantitative variables
5. `Labor Cost` - Quantitative variables
6. `Material Cost` - Quantitative variables
7. `State` - Qualitative variables

c. How many vehicles failed in month-9? (Write Excel function)

```
=COUNTIF(B:B, "9")
```

=71

d. What was the maximum labor cost? (Write Excel function)

```
=MAX(E:E)
```

=$3,234.41

e. What was the total failure cost for the data available? (Write Excel function)

`=SUM(E:F)`

=$685,836.05

Questions	Formulae	Answe
1.c. How many vehicles failed in month-9?	COUNTIF(B:B, "9")	71.0000
1.d. What was the maximum labor cost?	MAX(E:E)	$3,234.41
1.e. What was the total failure cost for the data available?	SUM(E:F)	$685,836.05

Problem-2: A seven-year medical research study reported that women whose mothers took the drug DES during pregnancy were *twice* as likely to develop tissue abnormalities that might lead to cancer as were women whose mothers did not take the drug.

a. This study involved the comparison of two populations. What were the populations?

- Women whose mothers took the drug DES during the pregnancy.
- Women whose mothers did not take the DES during the pregnancy.

b. Do you suppose the data were obtained in a survey or an experiment?

Since no attempt is made to control the variable of interest, data should be obtained by **survey**.

c. For the population of women whose mothers took the drug DES during pregnancy, a sample of 3980 women showed 63 developed tissue abnormalities that might lead to cancer. Provide a descriptive statistic that could be used to estimate the number of women out of 1000 in this population who have tissue abnormalities.

$$\text{Sample size} = 3980$$
$$\text{Number of women with tissue abnormalities} = 63$$
$$\text{n} = 1000$$
$$= \left(\frac{\text{Number of women with tissue abnormalities}}{\text{Sample size}}\right) \times n$$
$$= \left(\frac{63}{3980}\right) \times 1000 = 15.83 \approx 16$$

Estimates that approximately `16` out of 1000 women will have tissue abnormalities.

d. For the population of women whose mothers did not take the drug DES during pregnancy, what is the estimate of the number of women out of 1000 who would be expected to have tissue abnormalities?

We already found that approximately 15.83 out of 1000 women exposed to DES are expected to develop tissue abnormalities. The statement suggests that DES-exposed women are twice as likely to develop tissue abnormalities compared to those not exposed. To account for this, we can multiply the proportion of women with tissue abnormalities in the DES-exposed group by 0.5.

So, $15.83 \times 0.5 = 7.92 \approx 8$

Therefore, on average, approximately `8` out of 1000 women whose mothers did not take the drug DES during pregnancy are estimated to develop tissue abnormalities that might lead to cancer.

e. Medical studies often use a relatively large sample (in this case, 3980). Why?

The use of larger sample sizes in medical studies, like the 3980 participants in this case, serves a number of crucial purposes, such as producing more accurate, broadly applicable, and statistically significant results, which improves the research's credibility and dependability.

Problem - 3: ACNielsen conducts weekly surveys of television viewing throughout the United States. The ACNielsen statistical rating indicates the size of the viewing audience for each major network television program. Rankings of the television program and of the viewing audience market share for each network are published each week.

a. What is AC Nielsen attempting to measure?

AC Nielsen is attempting to measure the percentage of audience in the United States who watch each major network television program.

b. What is the population?

The population is all of the people in the United States who watch television.

c. Why would a sample be used in this situation?

AC Nielsen uses a sample because it is not possible to survey everyone in the United States who watches television.

d. What kinds of decisions or actions are based on the ACNielsen studies?

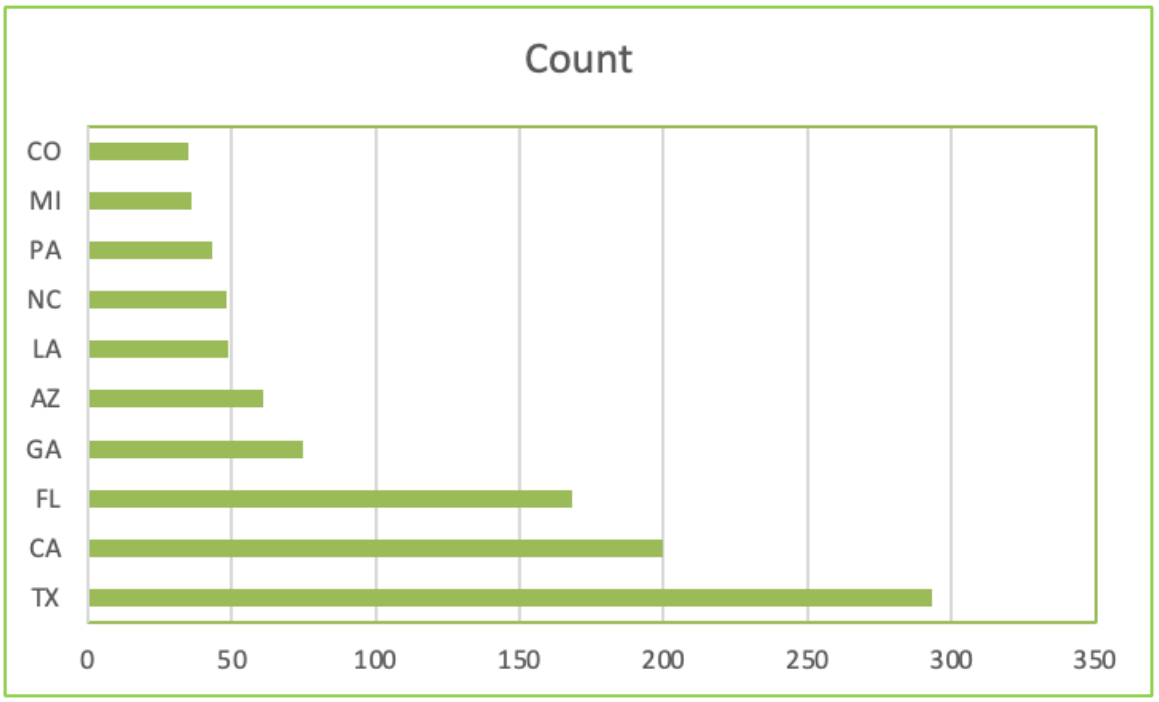
AC Nielsen studies play a pivotal role in guiding decisions across various domains, including aiding television regulators in comprehending program ratings, pinpointing the U.S. region with the most viewership, assisting advertisers in targeting specific audiences, and aiding television networks in determining show scheduling.

Problem - 4: Using 'VehicleFailureData', summarize the data for failures in top 10 (maximum number of vehicle failures) states by constructing the following:

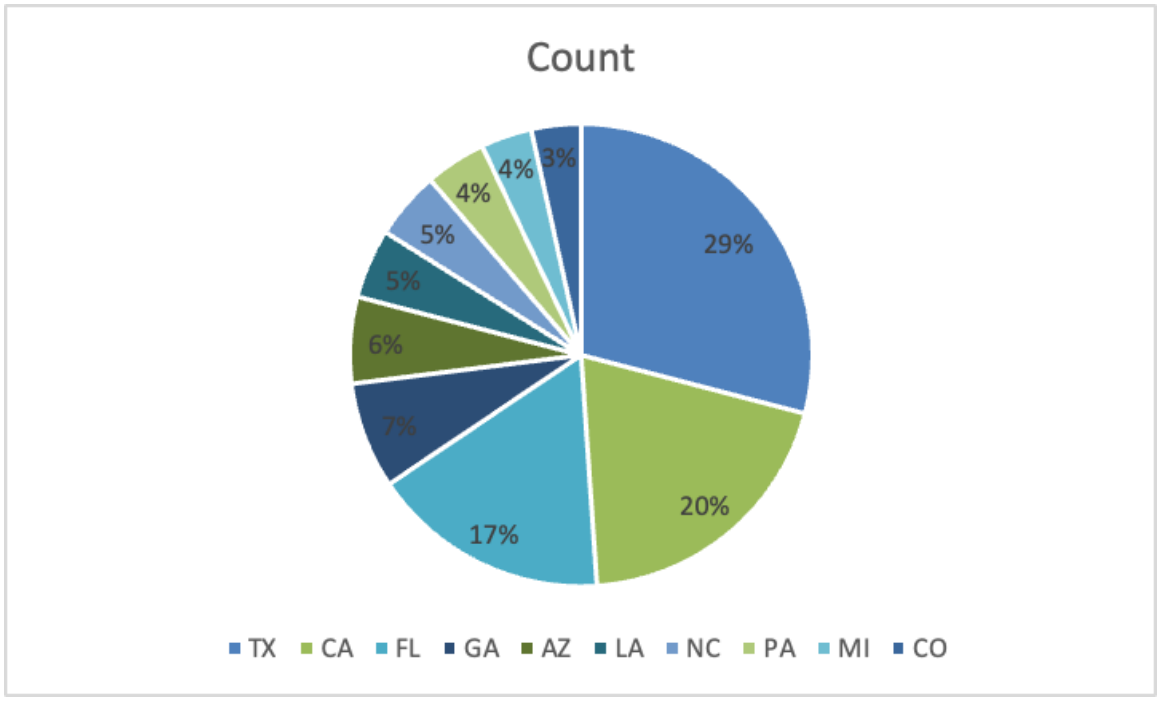
a. Relative and percent frequency distributions

State	Count	Relative	Percent
TX	293	0.29	29.07
CA	200	0.20	19.84
FL	168	0.17	16.67
GA	75	0.07	7.44
AZ	61	0.06	6.05
LA	49	0.05	4.86
NC	48	0.05	4.76
PA	43	0.04	4.27
MI	36	0.04	3.57
CO	35	0.03	3.47
Total	1008	1	100

b. Bar chart



c. Pie chart



d. Find top three states with maximum number of vehicles failures.

State	Count	Relative	Percent
TX	293	0.29	29.07
CA	200	0.20	19.84
FL	168	0.17	16.67

Case Study: Movie Theater Releases

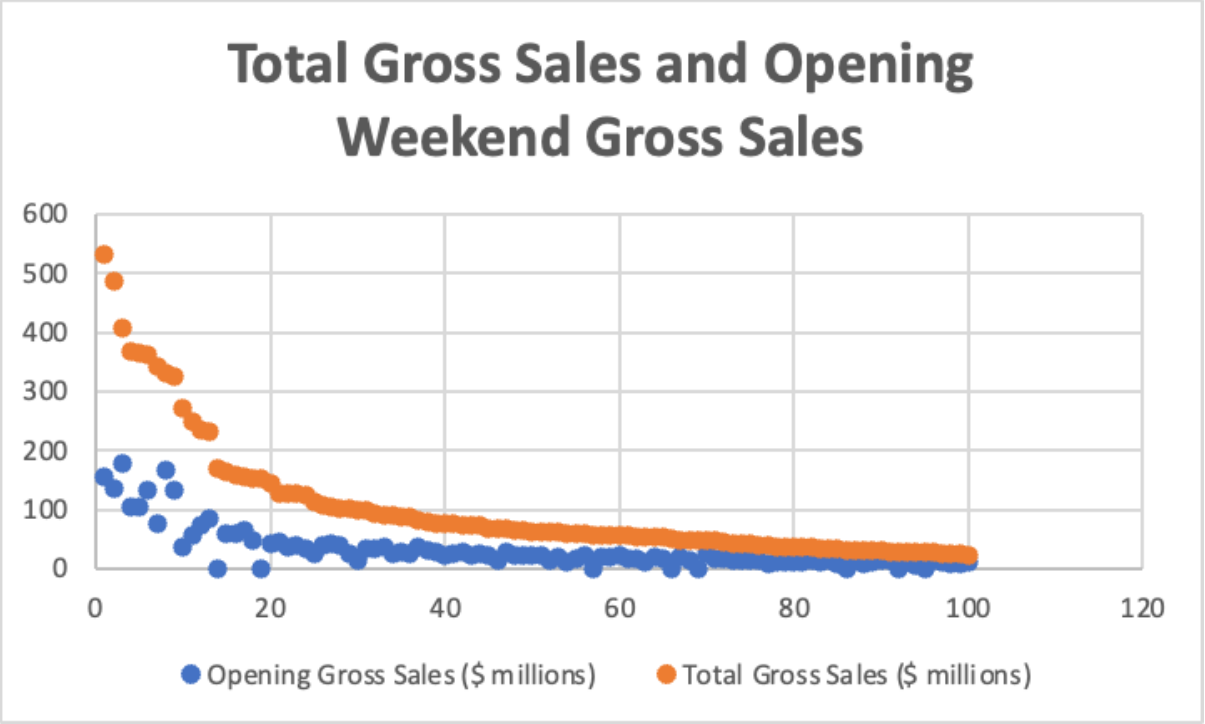
a. Tabular and graphical summaries for each of the four variables along with a discussion of what each summary tells us about the movies that are released to theaters.

Table Summary: This analysis explores the performance of 100 films from 2016, including metrics like Movie Title, Opening Gross Sales, Total Gross Sales, Number of Theatres, and Weeks in Release. Captain America: Civil War had the highest opening gross sales at $179.14 million, while Lion had the lowest at $0.12 million. Rogue One: A Star Wars Story led in total gross sales with $532.18 million, and Kevin Hart: What Now? had the lowest at $23.59 million. The Secret Life of Pets had the most theatres with 4381, while Manchester by the Sea had the fewest. Hidden Figures had the longest run with 46 weeks, and "Ben-Hur (2016)" ran for only 7 weeks.

Graphical Summary: To visually analyze the connection between opening weekend box office performance and overall box office success, a bar graph or scatter plot is recommended. A line graph can effectively illustrate the variation in the number of theatres allocated to each movie. Additionally, a bar graph can help visualize the correlation between the number of theatres and overall gross sales.

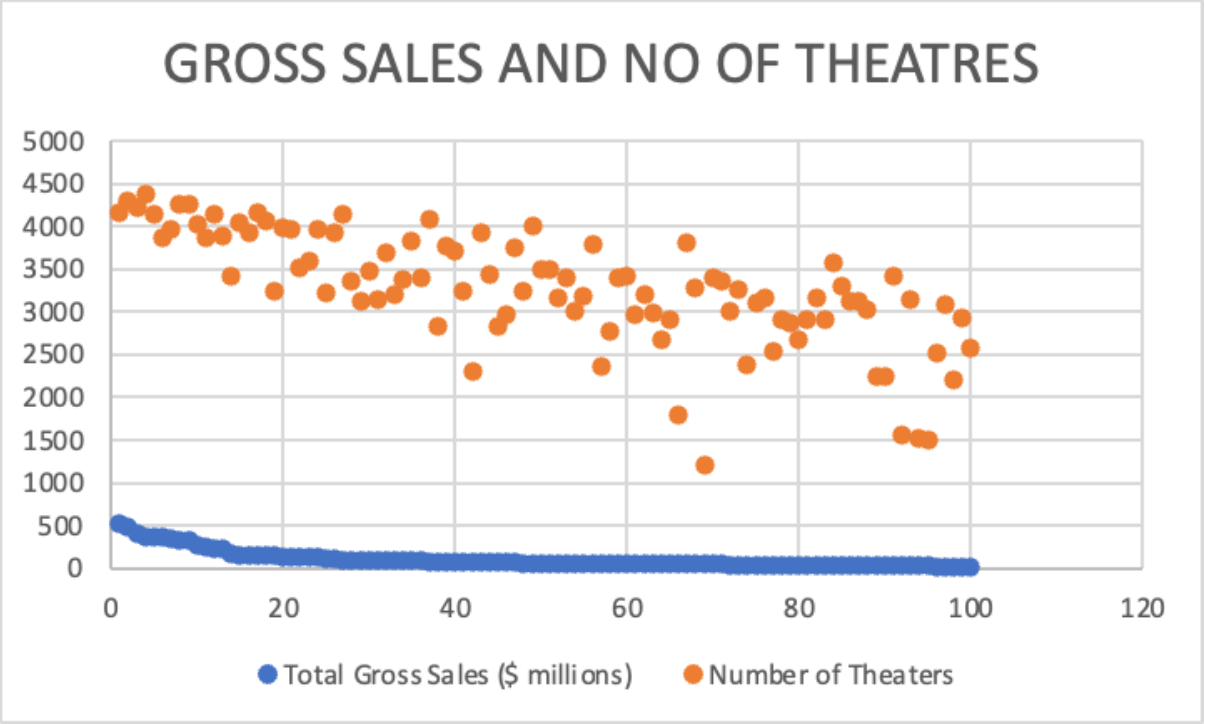
Discussion: The data and graphical summaries provide insights into the performance of 100 2016 films. Captain America: Civil War and Rogue One: A Star Wars Story were the top performers in both opening and total gross sales. Analyzing key factors like the number of theatres and weeks in release can reveal insights into a movie's success and distribution strategies. Films with more theatres and longer runs likely benefited from broader distribution and more extensive marketing efforts compared to those with fewer theatres and shorter runs.

b. A scatter diagram to explore the relationship between Total Gross Sales and Opening Weekend Gross Sales. Discuss.



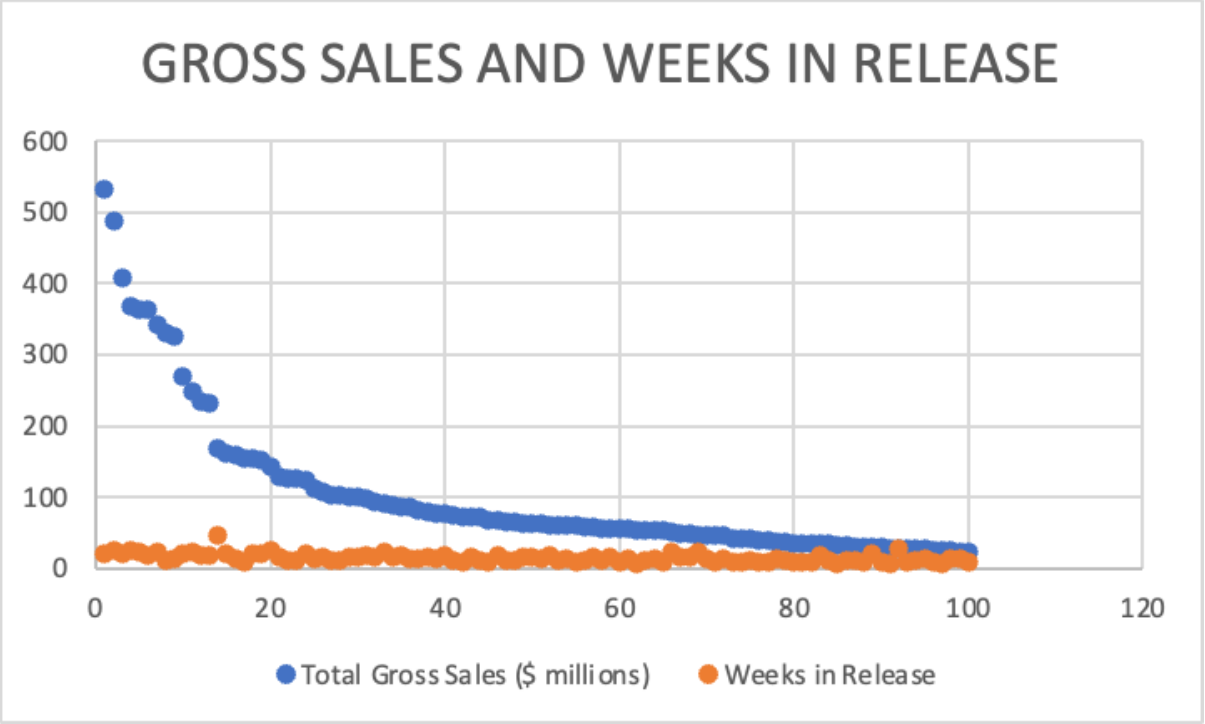
In the scatterplot above, we examine the relationship between total gross sales and opening gross sales for 100 films. The X-axis represents movie titles, and the Y-axis shows opening gross sales in millions of dollars. Captain America: Civil War had the highest opening weekend gross sales at $179.14 million, while "Lion" had the lowest at $0.12 million. Rogue One: A Star Wars Story achieved the highest overall gross sales at $532.18 million, with Kevin Hart: What Now? having the lowest at $23.59 million. Generally, there's a positive correlation between strong opening weekends and higher total gross sales. However, exceptions exist where films with weaker openings can still generate significant revenue due to positive reviews, word-of-mouth, and effective promotion. For instance, "Hidden Figures" had a modest opening of $0.52 million but accumulated an impressive total revenue of $169.61 million, highlighting the impact of sustained audience interest and positive reception over time.

c. A scatter diagram to explore the relationship between Total Gross Sales and Number of Theaters. Discuss.



The scatter diagram portrays the correlation between the number of theaters and total gross sales for 100 films. On the X-axis, movie titles are listed by their serial numbers using a scale of 20, while the Y-axis, scaled to 500, represents total gross sales. The orange line on the scatter plot represents the number of theaters for each film, while the blue line signifies total gross sales. Observations derived from the plot show that The Secret Life of Pets, with 4381 theaters, had the potential for extensive audience reach. However, Captain America: Civil War outperformed it in terms of total gross sales. Remarkably, Manchester by the Sea, despite being screened in the fewest theaters, achieved significant box office revenue, underscoring its strong performance compared to many other films with wider theater distributions.

d. A scatter diagram to explore the relationship between Total Gross Sales and Number of Weeks in Release. Discuss.



The scatter diagram illustrates how the number of weeks in release relates to total gross sales for 100 movies. The X-axis arranges movies by their serial numbers on a scale of 20, while the Y-axis depicts total gross sales on a scale of 100. In the scatter plot, the blue line represents total gross sales, and the orange line signifies the number of weeks in release for each movie. Several factors, such as critical reception, genre, cast, and competition from other films released concurrently, influence the correlation between total gross sales and weeks in release in this context. An intriguing observation emerges when we note that Hidden Figures, despite having the most weeks in release, did not achieve earnings as high as one might expect for movies with similar release durations. This observation suggests that additional factors played a role in shaping its box office performance.