Introduction to Diagnostic Analytics

Introduction to Analyzing Data for Business Goals

# What is Diagnostic Analytics?

As discussed in the previous modules, there are four types of data mining analytics as follows:

* **Descriptive:** What happened? Helps uncover valuable insight into the data being analyzed
* **Diagnostic:** Why did it happen? Helps understand the relationships and patterns in the data
* **Predictive:** What is likely to happen in the future? Helps forecast the future behavior of people and markets
* **Prescriptive:** What should I do about it? How should we respond to potential future events based on the analysis? Uses optimization and simulation algorithms to provide guidance and understanding on decisions and answers

This module introduces *diagnostic analytics*. Once we know *what* happened (thanks to descriptive analytics), we want to know *why* it happened. This is the purview of diagnostic analytics. For example, suppose December department store sales are lower this year than in previous years even though you expected them to be higher. You would turn to diagnostic analytics to help you drill down into the data and find out why this happened.

Diagnostic analytics helps you:

* Identify factors and potential causes from historical data
* Note relationships among data elements
* Decide where to “dig deeper” as you look for potential causes
* Answer questions posed by key stakeholders such as CEOs, CFOs, business strategists, and other high-level decision-makers in your organization

# Techniques of Diagnostic Analytics

Diagnostic analytics encompasses both *subjective* and *objective* approaches. Subjective analysis is based on data but also includes the opinions, interpretations, and judgment of scientists and engineers. Objective analysis, on the other hand, is purely fact-based, measurable, observable, and reproducible. Though seemingly in contradiction, these two approaches can be combined effectively to provide better guidance for analysts seeking explanations for observed phenomena.

There are three major types of diagnostic analytics:

**Identification of Inconsistencies**

Suppose you noticed that your company’s online sales increased significantly in one region of the country despite a lack of marketing efforts but *decreased* in a neighboring region that *did* receive more marketing attention. Diagnostic analytics can help find a reason for this seemingly inconsistent performance.

**Discovery** (drilling down into the data)

After you observe inconsistent behavior in a dataset, you may need to seek additional data that can help explain those observations. Diagnostic analytics can help you identify which supplementary data you need.

**Determining Underlying Relationships**

Discovering unexpected relationships among data elements may lead to a better understanding of the events that led to the observed inconsistencies. Probability theory, regression analysis, filtering, and time-series data analytics can all be useful for uncovering hidden stories in the data.

## Box Plots, Outliers, and Correlations

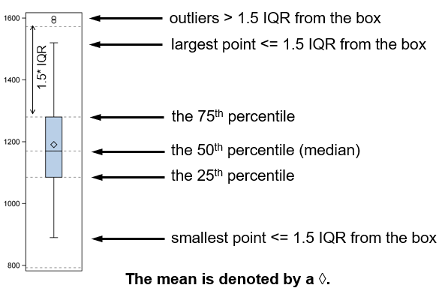
### What is a Box Plot?

In descriptive and diagnostic analytics, we often need more information about a dataset's distribution than just its measures of central tendency (median, mean, and mode). Figure 1 (below) shows how a *box plot* can be used to display the distribution of data based on five values:

* Minimum: Lowest value in the dataset (excluding outliers)
* 1st Quartile (*Q1* or *25th percentile*): Median value of the lower half of the dataset
* Median (*Q2* or *50th percentile*): Middle value of the dataset
* 3rd Quartile (*Q3* or *75th percentile*): Median value of the upper half of the dataset
* Maximum: Highest value in the dataset (excluding outliers)

Another important value is the *interquartile range* (IQR), which is defined as the difference between the 1st and 3rd quartiles.

IQR = Q3 – Q1



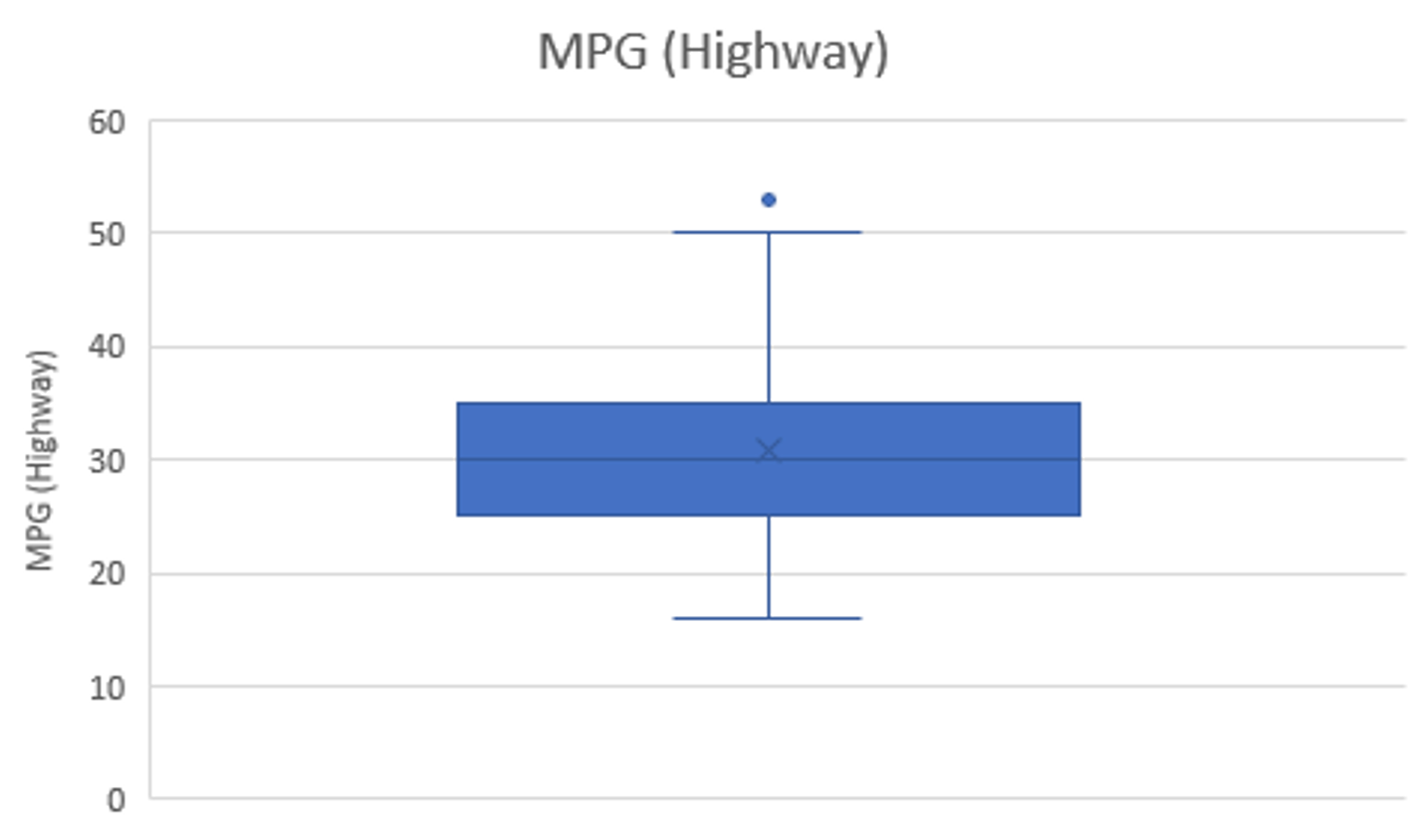
*Figure 1: An example of a box plot*

### What is an Outlier?

An *outlier* is any point in a dataset that is less than 1.5 IQR below Q1 or greater than 1.5 IQR above Q3.

Stated mathematically, the data point *n* is an outlier if **or**

As an example, consider a dataset containing automobile gasoline mileage data in miles per gallon (MPG) as summarized by the box plot in Figure 2. Note the outlier value depicted by a solid data point.

**

*Figure 2: Example of an outlier*

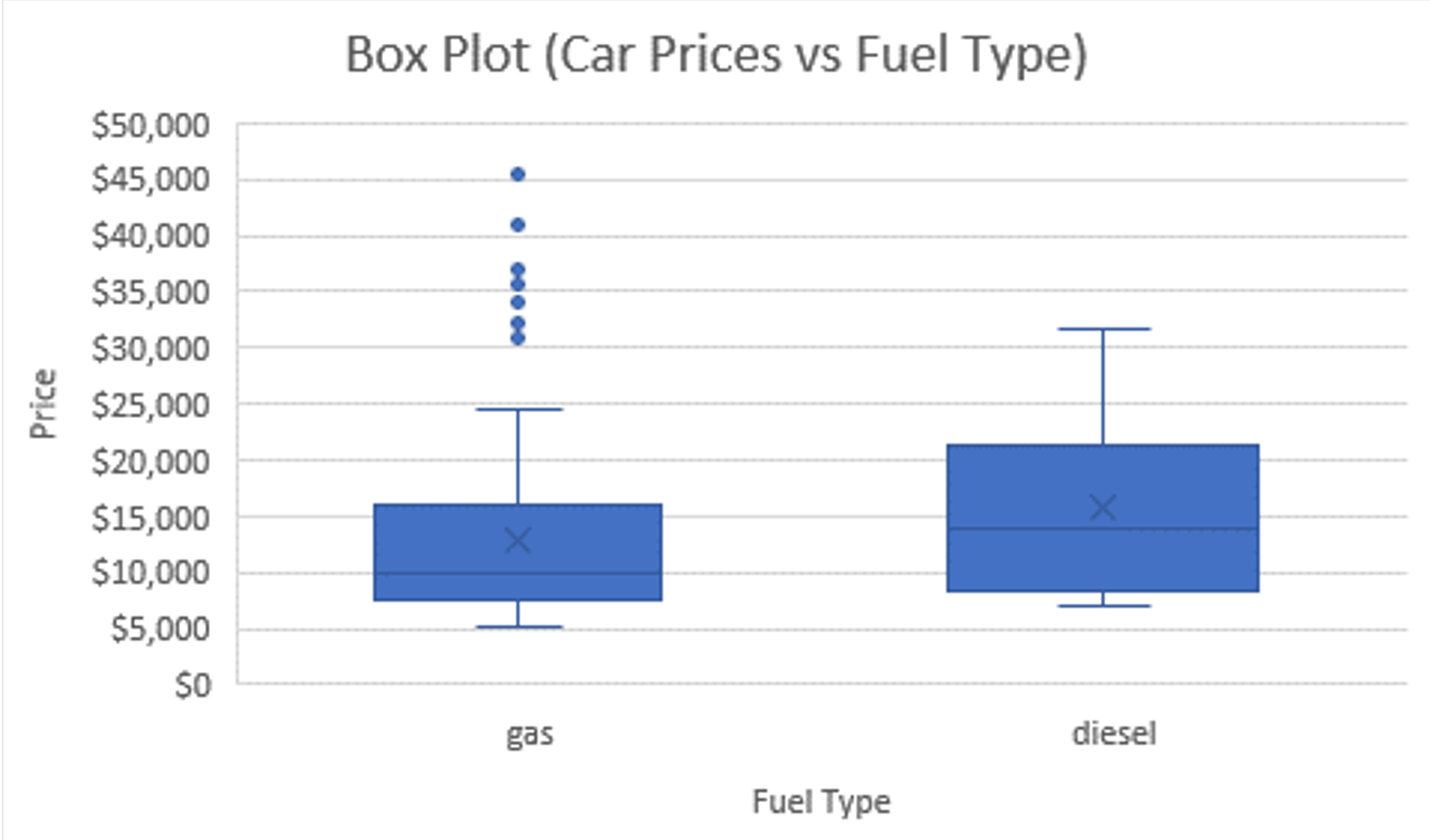
The mileage dataset has the descriptive statistics shown in Table 1.

| **Highway Mileage (miles per gallon)** | |
| --- | --- |
| Mean | 30.75122 |
| Standard Error | 0.48097 |
| Median | 30 |
| Mode | 25 |
| Standard Deviation | 6.886443 |
| Sample Variance | 47.4231 |
| Kurtosis | 0.44007 |
| Skewness | 0.539997 |
| Range | 38 |
| Minimum | 16 |
| Maximum | 54 |

*Table 1: Descriptive statistics for a dataset containing gasoline mileage data*

Note the outlier value of 54 MPG. The mean is approximately 31 MPG.

Now let’s compare the prices of gasoline-powered cars and diesel-powered cars (see the box plots in Figure 3).

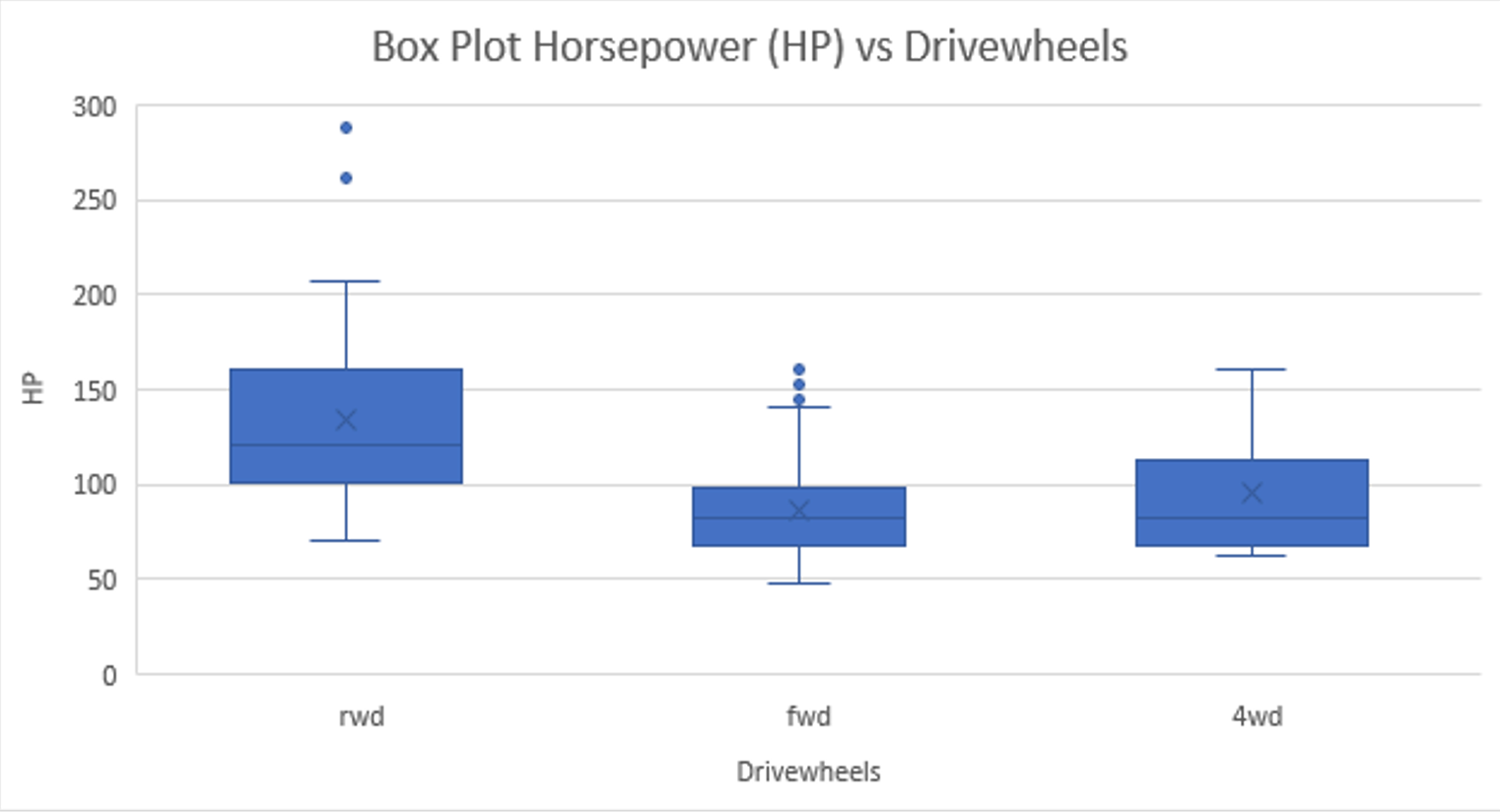


*Figure 3: Car prices—gas vs. diesel*

Here are several observations about the box plots in Figure 3.

* **Box Plot Heights:** The box plot for gasoline cars is shorter than the one for diesel cars. This tells you that the prices of gasoline cars cluster closer together than those for diesel cars. Generally, gasoline cars tend to cost about the same.
* **Box Plot Positions:** Note that the box plot for diesel cars is a bit higher than the one for gasoline cars. This suggests that diesel cars generally tend to be more expensive than gasoline cars.
* **Distribution Differences:** The median prices of gasoline cars and diesel cars are not very far apart—$12,916 for gasoline cars and $15,838 for diesel cars. However, the box plots show that the distribution of data points is much wider for diesel cars than it is for gasoline cars. Looking at only the medians, you might conclude that the two kinds of automobile are not much different in terms of price. However, expanding your view and looking at the overall distribution of the data points, you’ll conclude that there are potentially some important differences after all.
* **Outliers:** As you can see, there are quite a few outliers among the gasoline car data points: they range from $30,760 to $45,400.

Continuing our automobile data exploration, Figure 4 compares automobile horsepower (HP) for read-wheel drive (rwd), front-wheel drive (fwd), and four-wheel drive (4wd) vehicles.



*Figure 4: A comparison of automobile horsepower for three different drivetrains*

Interestingly, we see right away that rear-wheel drive vehicles have the two highest outliers (262 and 288 hp). The median is 133 hp.

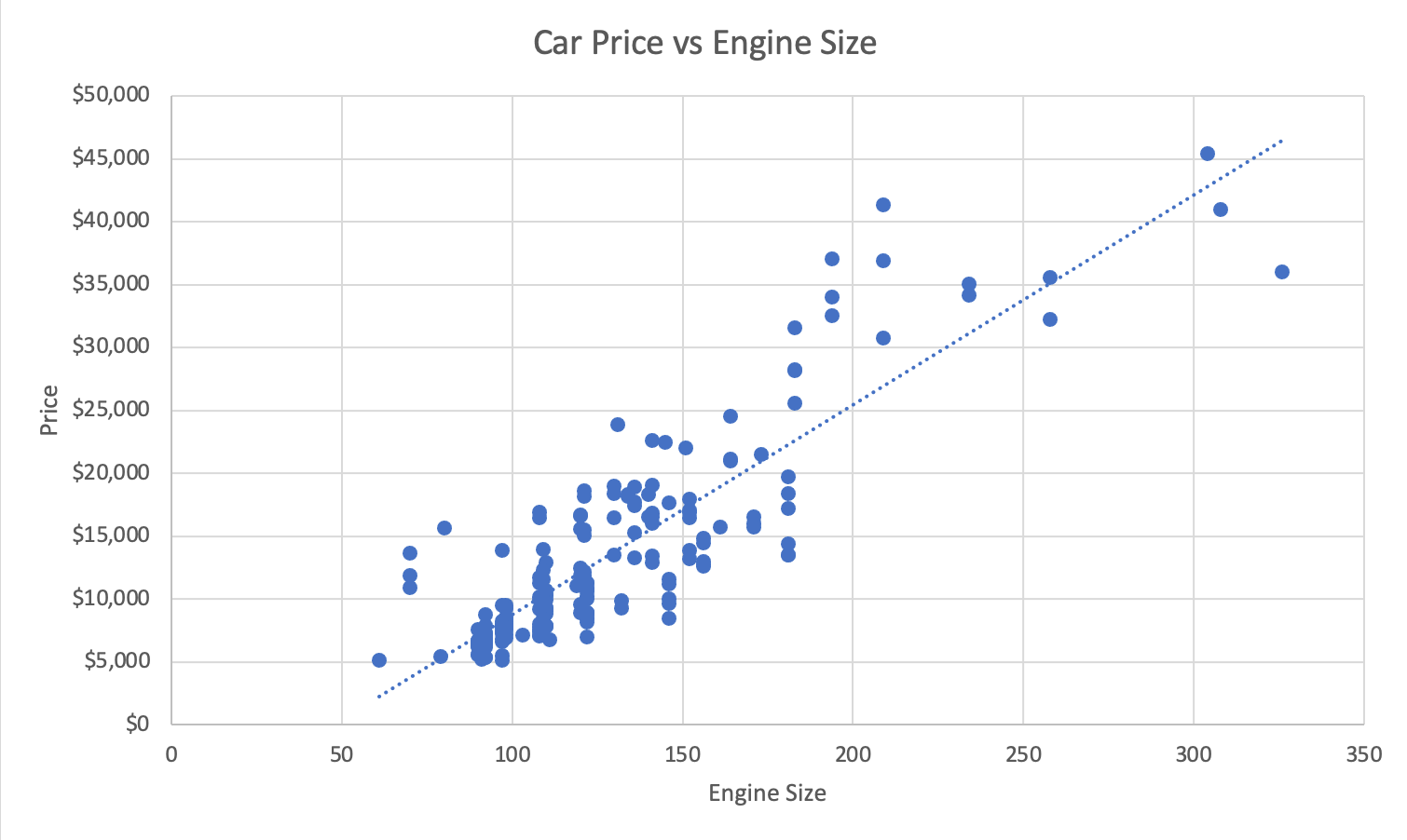
As you can see, box plots provide a large amount of information in a very concise manner. They are especially useful as you begin your data analysis because they can help identify patterns and outliers early in the process.

### What is a Correlation?

In descriptive and diagnostic analytics, we seek to identify patterns in the data. One way to do this is to use a scatter plot and look for correlations.

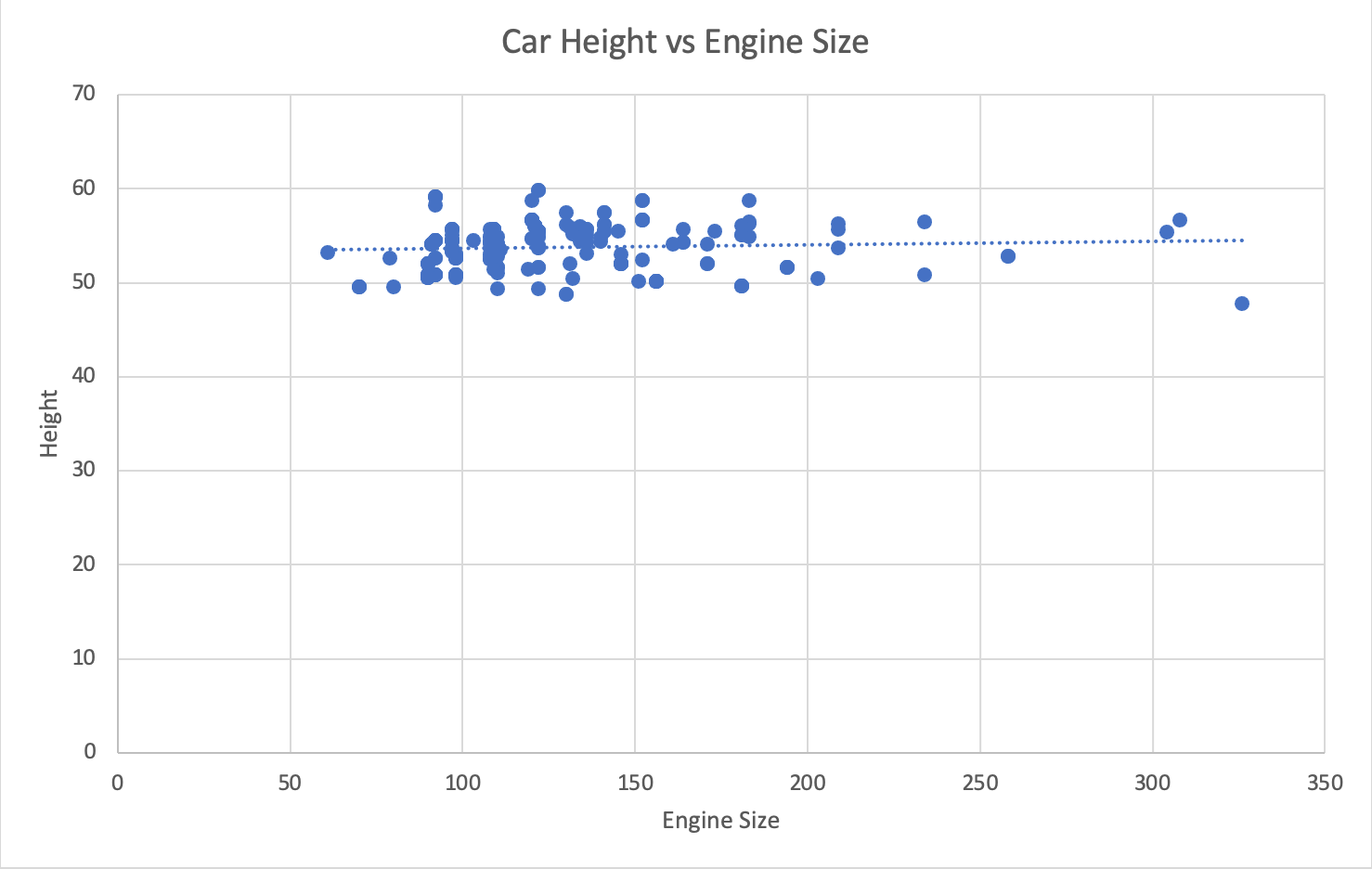
A scatter plot is a graph of a bivariate distribution—two variables (typically labeled as *x* and *y*) that are paired with each other. Naturally, you would plot two variables in this manner because there is a logical connection or relationship between the two.

For example, let’s look at the relationship between car prices and engine size. To generate such a plot, you take price and engine size data for each car, and plot them on a graph (Figure 5). This graph readily shows that cars with bigger engines tend to cost more. This makes sense because it’s reasonable that larger engines are more expensive to manufacture. As the scatter plot shows, the relationship between price and engine size is positive—that is, a bigger engine means a higher price. (Higher values for one variable correspond to higher values for the other variable.)



*Figure 5: A scatter plot showing car prices with respect to engine size*

In contrast, plotting car heights with respect to engine size shows that all cars generally have the same height regardless of engine size (Figure 6). This seems reasonable since all cars—even those with large engines—need to fit under bridges and inside people’s garages. If you know the height of a particular car, you can’t really say much about the size of its engine.



*Figure 6: A scatter plot showing car heights with respect to engine size*

*Correlation* is a numerical measure of the strength and direction of the linear relationship between two continuous variables (*x* and *y*). While a scatter plot gives us a visual idea of how the two variables are related, the correlation gives us a numerical value—the *correlation coefficient*—which provides a quantitative measure of the dataset’s linearity. The correlation coefficient, denoted as *r*, ranges from –1 to +1, with a positive value indicating a positive association and a negative value indicating a negative association. If *r* is close to zero, the two variables are only weakly associated. (Remember, this discussion pertains specifically to *linear* relationships.)

The correlation coefficients for pairs of variables in the automobile dataset were calculated and are presented in Table 2. Note that the weakest relationship is between car heights and engine sizes (r = 0.067) and the strongest relationship is between car price and engine size (r = 0.872).

Figure 7 (two pages ahead), generated by Tableau, shows several of the correlations depicted graphically, which allows you to confirm visually whether the correlations are strong, weak, or nonexistent.

In broader terms, the correlation of two random variables *x* and *y* reflects the strength of the linear relationship between them. If *x* and *y* are positively correlated, then we can feel reasonably confident that observing a large value of *x* will likely mean that the value of *y* is also large.

|  | **Wheel base** | **Length** | **Width** | **Height** | **Engine**  **size** | **HP** | **City mpg** | **Hwy**  **mpg** | **Price** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Wheelbase | 1 |  |  |  |  |  |  |  |  |
| Length | 0.8745 | 1 |  |  |  |  |  |  |  |
| Width | 0.7951 | 0.8411 | 1 |  |  |  |  |  |  |
| Height | 0.5894 | 0.4910 | 0.2792 | 1 |  |  |  |  |  |
| Engine Size | 0.5693 | 0.6833 | 0.7354 | **0.0671** | 1 |  |  |  |  |
| Hp | 0.3522 | 0.5550 | 0.6424 | -0.1107 | 0.8107 | 1 |  |  |  |
| City  mpg | -0.4704 | -0.6709 | -0.6427 | -0.0486 | -0.6536 | -0.8036 | 1 |  |  |
| Hwy  mpg | -0.5440 | -0.7046 | -0.6772 | -0.1073 | -0.6774 | -0.7709 | 0.9713 | 1 |  |
| Price | 0.5846 | 0.6906 | 0.7512 | 0.1354 | **0.8723** | .08105 | -0.6865 | -0.7046 | 1 |

*Table 2: Correlations between pairs of automobile data variables (truncated to five significant figures)*

## Sensitivity Analysis

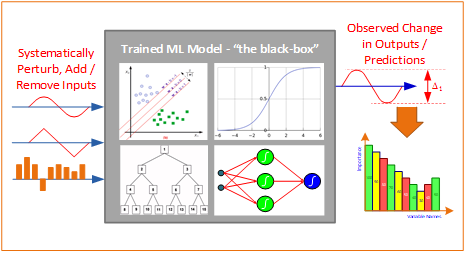
Ways of answering the question, “Why did it happen?” are not limited to traditional statistical methods like correlation analysis, outlier identification, and regression modeling. Contemporary methods also include *sensitivity analysis*, which is a family of variable explanation techniques often applied to machine learning models. Sensitivity analysis is also called *explainable AI* (XAI), *interpretable ML*, or *lightening the “black-box”* (where the black-box is a metaphor for the ML model). The goal of sensitivity analysis is to identify the input variables that are most important for the determination of the output variable. That is, if you know which variables play the more significant roles in determining the value of the target variable (e.g., customer churn), you can deduce the underlying reasons and directly (or indirectly) answer the question of why some of your customers are leaving (i.e., churning), and what you can do to potentially prevent them from leaving.

## 

*Figure 7: Correlations represented graphically*

Figure 8 graphically illustrates the sensitivity analysis process (Delen, 2020). In the middle, we have four different types of machine learning techniques. On the left side we have the input variables to these models. And on the right side we have the output variable. There are two main approaches we can follow:

1. Systematically perturb each of the input variables individually (within an appropriate range) while keeping the other variables constant, and observe changes in the output variable. The input variable that causes the greatest output variable change is thus the most important or significant variable. After doing this for each input variable, you’ll have a list of variables ranked in order of importance, which you can use for developing a machine model.
2. Systematically omit each of the input variables and observe the effects. This *leave-one-out* approach, as it’s called, is unlike the perturbation approach in that you omit variables entirely rather than perturb them. First, develop a baseline prediction model that includes all the input variables. Second, leave out one of the input variables and develop the same model using the remaining input variables. Assess this model’s prediction accuracy and calculate the drop in its predictive value (i.e., the difference in prediction accuracy between the baseline model and the new model). Third, repeat the process by repeatedly leaving out each of the remaining variables and noting the reductions in predictive accuracy when each variable is omitted. Fourth, rank-order the input variables according to the drops in prediction accuracy. The most important variable is the one for which the prediction accuracy drops the most when it was omitted.



*Figure 8: A graphical depiction of the sensitivity analysis process*

# Achieving Business Goals with Diagnostic Analytics

Needless to say, there is tremendous business value in discovering why certain things happen. Knowing and thoroughly understanding the underlying causes of operational problems or off-target performance results helps business leaders address those problems, set better goals, and develop more effective strategies for improving performance. As a data analyst, it’s your job to conduct the analyses and interpret the results so that your stakeholders can make informed decisions.

As discussed in previous modules, there are a variety of ways to present data and analytical results to your audience using tables, charts, graphs, diagrams, and so forth. You then need to interpret these visual representations to provide both context and meaning that your audience can convert into action. The process of diagnostic analytics is best illustrated with an example.

## Diagnostic Analytics Example

You’re the assistant manager of a coffee shop in a mid-sized college town. Your shop is typically busy seven days per week with a broad mixture of clients including college students, people on their way to work (and on their way home), and families stopping by on weekends. Sales revenue is about the same regardless of the day of the week.

Your new boss, who was recently hired as manager, is trying to understand the shop’s sales patterns. He shows you this table of monthly sales figures:

| **Month** | **Sales** |
| --- | --- |
| Apr | $ 121,467 |
| May | $ 110,498 |
| Jun | $ 113,708 |
| Jul | $ 120,017 |
| Aug | $ 129,450 |
| Sep | $ 119,451 |
| Oct | $ 104,406 |
| Nov | $ 99,895 |
| Dec | $ 94,384 |
| Jan | $ 125,214 |
| Feb | $ 128,509 |
| Mar | $ 96,239 |
| Apr | $ 117,534 |
| ***Total*** | ***$ 1,480,772*** |

*Table 1: Monthly sales*

He understands that sales might be lower in November and December because of the holidays as people go out of town, but he is wondering why March isn’t higher and asks you to explore further.

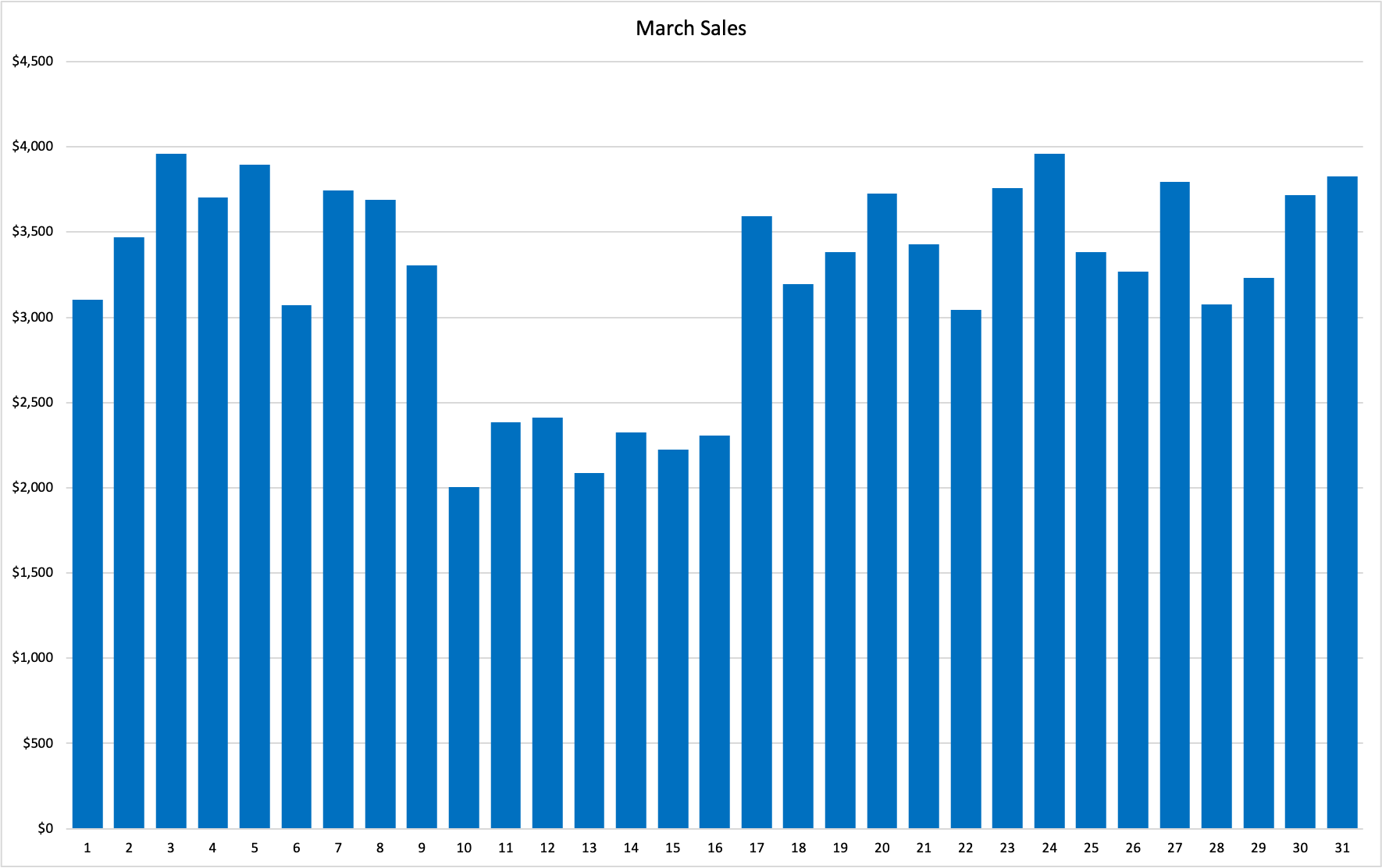
The first thing you do is look at daily sales for March (Table 2).

| Day | March Sales |
| --- | --- |
| 1 | $ 3,101 |
| 2 | $ 3,471 |
| 3 | $ 3,958 |
| 4 | $ 3,703 |
| 5 | $ 3,895 |
| 6 | $ 3,072 |
| 7 | $ 3,746 |
| 8 | $ 3,691 |
| 9 | $ 3,304 |
| 10 | $ 2,003 |
| 11 | $ 2,384 |
| 12 | $ 2,413 |
| 13 | $ 2,088 |
| 14 | $ 2,325 |
| 15 | $ 2,222 |
| 16 | $ 2,307 |
| 17 | $ 3,595 |
| 18 | $ 3,195 |
| 19 | $ 3,384 |

| 20 | $ 3,724 |
| --- | --- |
| 21 | $ 3,427 |
| 22 | $ 3,044 |
| 23 | $ 3,759 |
| 24 | $ 3,959 |
| 25 | $ 3,384 |
| 26 | $ 3,267 |
| 27 | $ 3,793 |
| 28 | $ 3,077 |
| 29 | $ 3,230 |
| 30 | $ 3,717 |
| 31 | $ 3,826 |

*Table 2: Daily sales for March*

The numbers vary as you would expect, but no obvious pattern is immediately visible. So, you plot the values (Figure 7).



*Figure 7: March sales data*

Now you can easily see that something unusual happened during the week beginning on March 10. You show this chart to one of the coffee shop employees (who happens to be a doctoral student in statistics earning some extra money on the side) and she casually remarks, “Oh, that was spring break week. The campus was closed.”

That’s it—there were many fewer student customers that week.

This relatively simple example illustrates the basic ideas behind diagnostic analytics. Someone observes something they can’t explain so they gather some data and analyze it. Very often, a visual representation helps uncover a pattern—or a break in a pattern—which then points to something specific that merits further questioning and investigation. In more complex situations, you might have to apply statistical analyses involving modeling and hypothesis testing, but the basic idea is the same: you strive to answer the question, “Why did this happen?” and then present the answer to stakeholders for further action.

In the example, you have a nice chart to show the manager and an explanation to offer. The manager can decide to stop worrying about the drop in sales knowing that it’s only temporary or perhaps plan a promotional event for the following year’s spring break week in an effort to increase non-student business.

# References

* Delen, D. (2021). Predictive Analytics: Data Mining, Machine Learning and Data Science for Practitioners 2nd Edition. Upper Saddle River, New Jersey: Pearson (FT Press Analytics), 978-0136738510.
* Vesset, D. (2018). [IBM: Descriptive analytics 101: What happened?](https://www.ibm.com/blogs/business-analytics/descriptive-analytics-101-what-happened/)