Sec 4.11 – Using Mathematical Models
Big Data

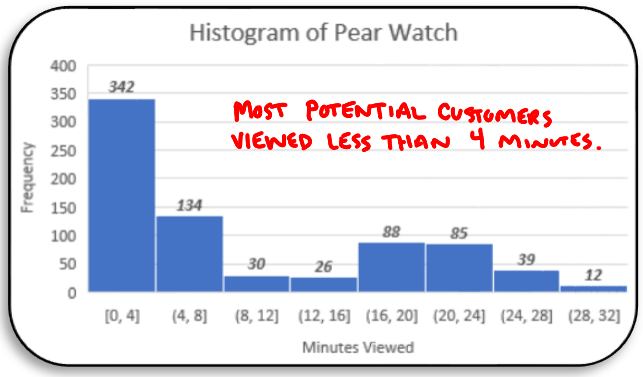
Name:

"Big Data" is considered to be large, diverse, data sets that can be rapidly obtained. Often big data is referred to by the three **V**'s: a large **Volume** of data, a **Variety** of data information, and a high **Velocity** at which data is obtained (*sometimes Value and Veracity are also included*).

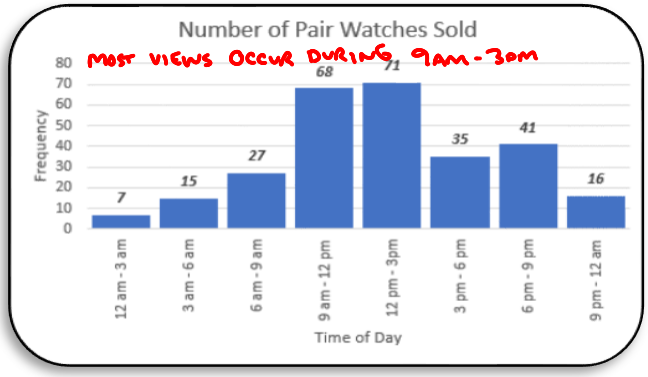
Consider an example of an online store like Amazon investigating big data about a new Smart Watch they are selling the Pear Watch.

Amazon first investigated **structured quantitative big data** by monitoring a ***variety*** of topics ***quickly***.

- They monitored the time of each potential customer spent looking at a new Smart Watch product web page on their site for the initial launch day.

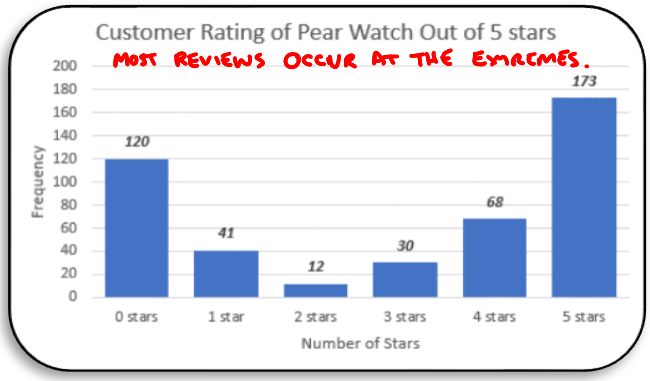


- They monitored the peak times at which they sold the Smart Watch.



- Over a period of weeks, Amazon experimented with moderately varying the price of watch to determine how much the price impacted sales.

- They investigated the distribution of the customer ratings of the product.



Amazon also investigated **unstructured qualitative big data**.

- They conducted a web search analysis of social media to find comments made about the smart watch. The report just listed the comments.

- They conducted an analysis on what key words were used most often when searching for the new Smart Watch. The report listed words used when searching in Google for information about a Pair Watch.

Describe a statement you could make about each one of these potential data sets and distributions.

- **Data Integration** is the process of formatting and combining data from multiple sources into a single unified view so that it can be more readily statistically analyzed.

- **Data Mining** is the process of examining large and varied data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences, and other useful information.

- **Data Warehousing** is the process of storing and managing data in a way that allows for efficient retrieval and analysis.

The relationship between 2 variable interest is referred to as the **correlation**. The correlation can be described by many different models such as linear, quadratic, exponential, logistics, etc. To compare how well that various model fit the 2 variables is measured by the Goodness-of-Fit analysis usually referred to as the **coefficient of determination (R^2).** The closer the R^2 value is to 1 the better the model is at suggesting the type of relationship between the 2 variables of interest.

Let's look at some practical applications of big data. Amazon knows that consumers sometimes associate price with quality (i.e. the more expensive an item is the better they think it is). This is more likely to be true when consumers purchase generic items. So, they collect data on the number of an item sold at various prices. Let's look at some data on the exact same generic Pulse Oximeter that Amazon tried selling at various prices across the U.S.

Price	$9.99	$17.89	$24.59	$29.99	$35.89	$46.99
Units Sold	658	2,492	3,101	3,024	2,631	781

1. Create a Scatter Plot.

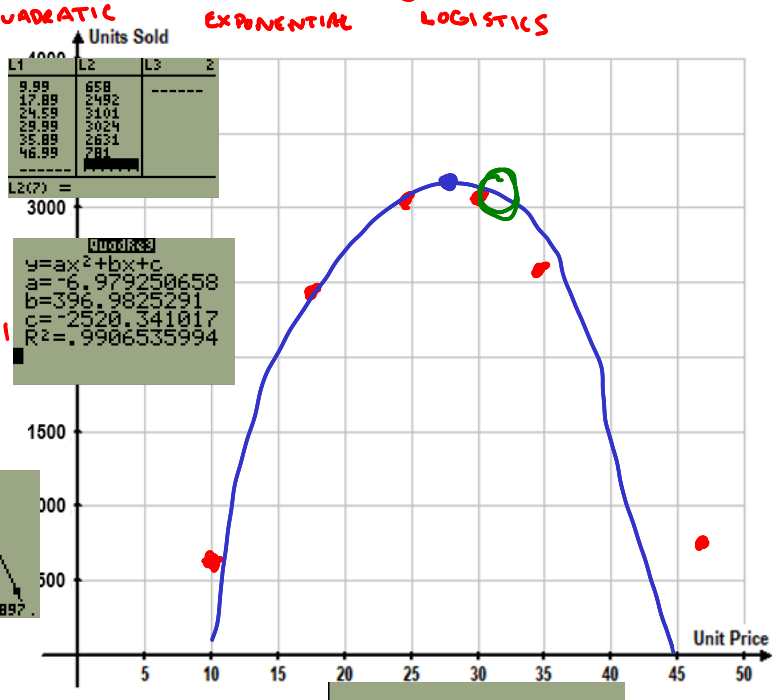
LINEAR QUADRATIC EXPONENTIAL LOGISTICS

2. What type of mathematical model would you suggest using to describe the type of correlation?

QUADRATIC

3. Use your Graphing Calculator to determine a regression equation that models number of units sold as a function of price.

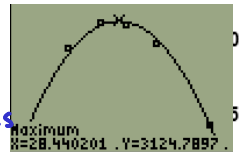
$Y = -6.979x^2 + 396.983x - 2520.341$



4. Using your mathematical model, what price would you recommend listing to maximize sales?

$28.44 ← SELL MOST

FIND WHERE THE MAXIMUM OF $Y = -6.979x^2 + 396.983x - 2520.341$ OCCURS



5. Using your mathematical model what price would you recommend listing to maximize gross profit?

$34.42 ← MOST PROFIT

FIND WHERE THE LOCAL MAX OF

$Y = X \cdot (-6.979x^2 + 396.983x - 2520.341)$ OCCURS.