



IIT School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

information technology & management

527 Data Analytics

January 26 2016

Week 3 Presentation

Week 3 Topic: Expectation & Agenda

- ◆ Basics of Statistics
- ◆ Reviewed course materials from last week - Optimization Modeling
- ◆ Case Study – Pothole Repair

Week 3 Topic:

Parameters and Statistics

- ◆ Statistics are used to approximate population parameters.
- ◆ *Parameters* are characteristics of populations. Because populations usually cannot be measured in their entirety, parameter values are generally unknown. *Statistics* are quantities calculated from the values in the sample.

| | Population Parameters | Sample Statistics |
|--------------------|-----------------------|-------------------|
| Mean | μ | \bar{x} |
| Variance | σ^2 | s^2 |
| Standard Deviation | σ | s |

Week 3 Topic:

Statistics

Descriptive Statistics:

- ◆ The goals when you are describing data are to
 - screen for unusual sample data values
 - inspect the spread and shape of continuous variables
 - characterize the central tendency of the sample.

Inferential Statistics:

- ◆ The goals for statistical inference are to
 - estimate or predict unknown parameter values from a population, using a sample
 - make probabilistic statements about population attributes.
- ◆ After you select a random sample of the data, you can start describing the data. Although you want to draw conclusions about your population, you first want to explore and describe your data before you use inferential statistics. Why?
 - Data must be as error free as possible.
 - Unique aspects, such as data values that cluster or show some unusual shape, must be identified.
 - An extreme value of a variable, if not detected, could cause gross errors in the interpretation of the statistics.

Week 3 Topic:

Parameters and Statistics

- ◆ *Statistics* are quantities calculated from the values in the population.
- ◆ Suppose you have x_1, x_2, \dots, x_n , a sample from some population

$$\bar{x} = \frac{1}{n} \sum x_i$$

The mean is an average, a typical value in the distribution.

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

The variance measures the sample variability.

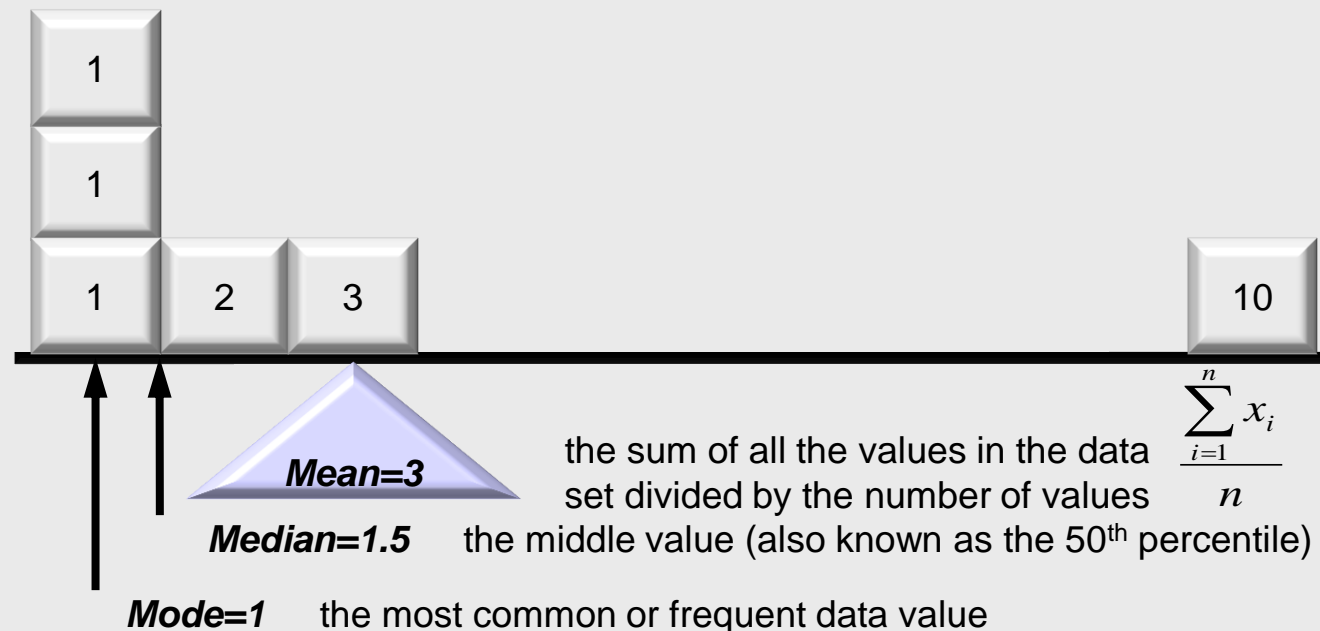
$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

The standard deviation is square root of the variance and also measures variability in data. Usually reported in the same units as the mean.

Week 3 Topic:

Mean, Median, Mode

- ◆ A property of the sample mean is that the sum of the differences of each data value from the mean is always 0, that is, $\sum (x_i - \bar{x}) = 0$.
- ◆ The *mean* is the arithmetic balancing point of your data.
- ◆ The *median* is the data point in the middle of a sorted sequence. It is appropriate for either rank scores (variables measured on an ordinal scale) or variables measured on an interval or ratio scale with a skewed distribution.
- ◆ The *mode* is the data point that occurs most frequently. It is most appropriate for variables measured on a nominal scale. There might be several modes in a distribution.



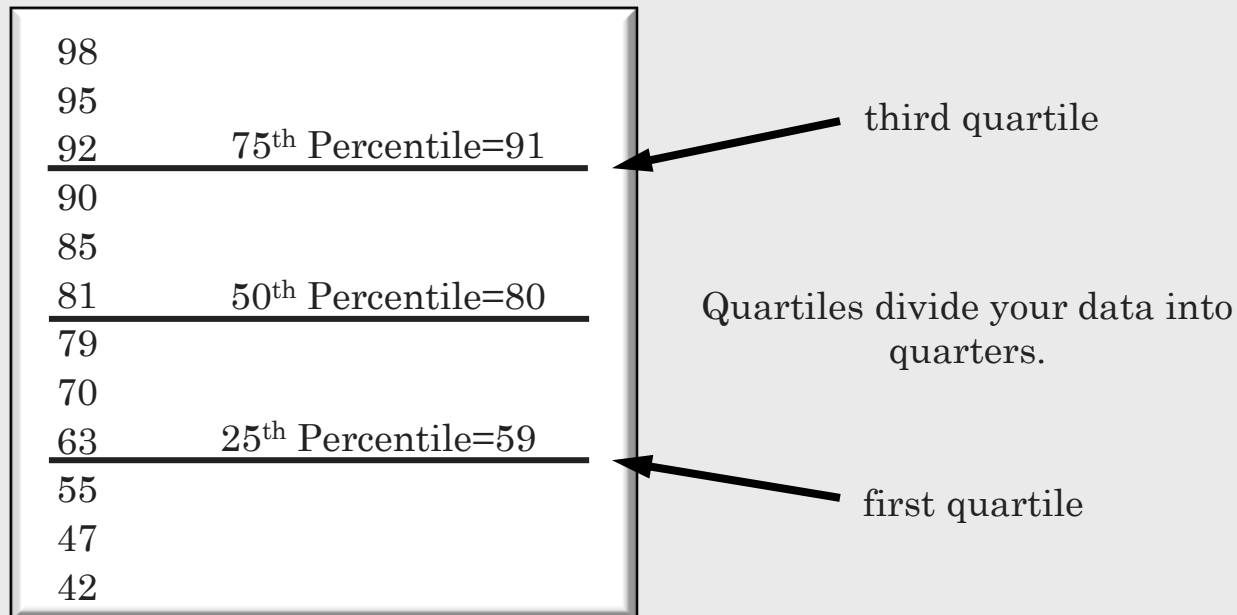
Week 3 Topic:

Distributions

- ◆ A *distribution* is a collection of data values that are arranged in order, along with the relative frequency. For any type of data, it is important that you describe the location, spread, and shape of your distribution using graphical techniques and descriptive statistics.
- ◆ When you examine the distribution of values in a variable, you can determine the following:
 - the range of possible data values
 - the frequency of data values
 - whether the data values accumulate in the middle of the distribution or at one end
 - Are the values symmetrically distributed?
 - Are any values unusual?
 - What is the best estimate of the average of the values for the population?
 - What is the best estimate of the average spread or dispersion of the values for the population?

Week 3 Topic: Percentiles

- ◆ *Percentiles* locate a position in your data larger than a given proportion of data values.
- ◆ These are commonly reported percentile values:
 - the 25th percentile, also called the first quartile
 - the 50th percentile, also called the median
 - the 75th percentile, also called the third quartile



Week 3 Topic:

Spread of Distribution - Dispersion

- ◆ Measures of dispersion enable you to characterize the dispersion, or spread, of the data.
- ◆ A value better suited to reflect dispersion is the *interquartile range*. The interquartile range shows the range of the middle 50% of data values.

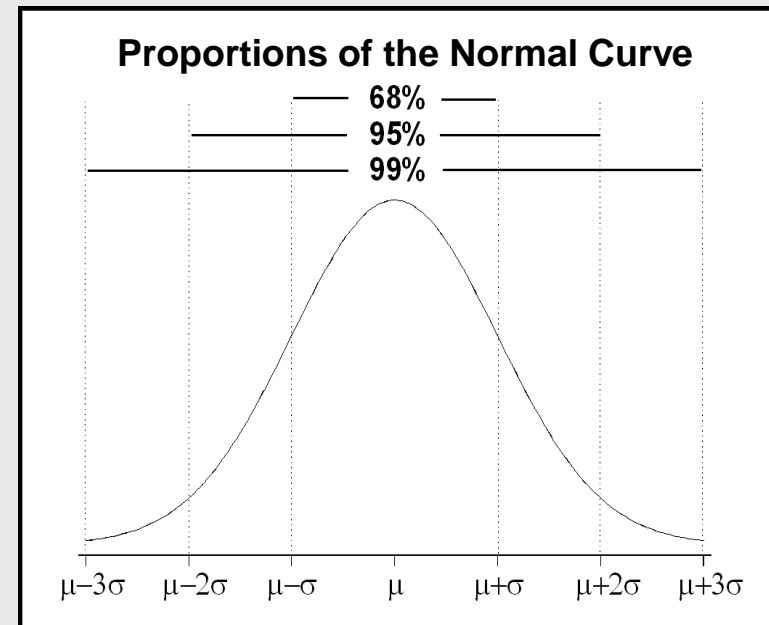
| Measure | Definition |
|---------------------|---|
| Range | the difference between the maximum and minimum data values |
| Interquartile Range | the difference between the 25th and 75th percentiles |
| Variance | a measure of dispersion of the data around the mean |
| Standard Deviation | a measure of dispersion expressed in the same units of measurement as your data (the square root of the variance) |

Week 3 Topic:

Normal Distribution

A normal distribution

- ◆ is symmetric. If you draw a line down the center, you get the same shape on either side.
- ◆ defined by two parameters, μ (the population mean) and σ (the population standard deviation).
- ◆ is bell shaped and has mean = median = mode which describes the midpoint of the distribution
- ◆ Another name for the normal distribution is the Gaussian distribution.
- ◆ The standard normal curve has $\mu=0$ and $\sigma=1$. The area under the curve between any two values can be calculated.
- ◆ Approximately 68% of the total area lies within 1 standard deviation of the mean.
- ◆ Approximately 95% of the total area lies within 1.96 standard deviations of the mean.
- ◆ Approximately 99.7% of the area lies within 3 standard deviations of the mean.



Week 3 Topic:

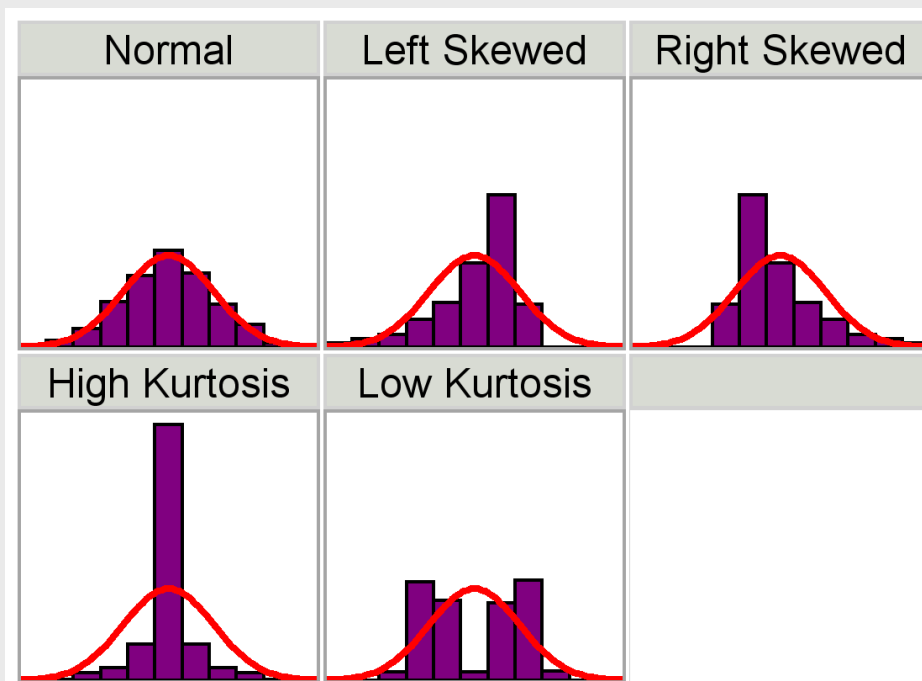
Normal Distribution (cont.)

- ◆ Often in analysis, although not always, a normal distribution is assumed.
- ◆ The normal distribution is a mathematical function. The height of the function at any point on the horizontal axis is the “probability density” at that point. Normal distribution probabilities (which can be thought of as the proportion of the area under the curve) tend to be higher near the middle.
- ◆ The center of the distribution is the population mean (μ). The standard deviation (σ) describes how variable the distribution is about μ . A larger standard deviation implies a wider normal distribution. The mean locates the distribution (sets its center point) and the standard deviation scales it.
- ◆ Often, values that are more than two standard deviations from the mean are regarded as unusual. Only about 5% of all values are at least that far away from the mean.
- ◆ You use this information later when you discuss the concepts of confidence intervals.

Week 3 Topic:

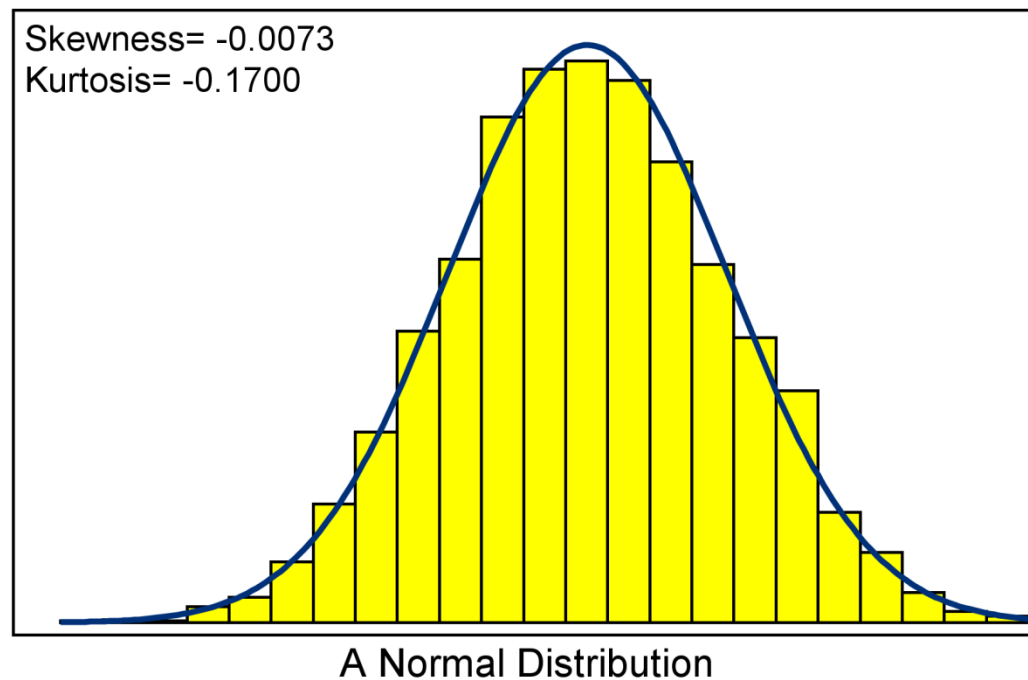
Distributions compared to Normal

- ◆ The distribution of your data might not look normal. There are an infinite number of ways that a population can be distributed. When you look at your data, you might notice the features of the distribution that indicate similarity or difference from the normal distribution.
- ◆ When you evaluate distributions, it is useful to look at statistical measures of the shape of the sample distribution compared to the normal.
- ◆ Two such measures are skewness and kurtosis.



Week 3 Topic: Normal Distribution

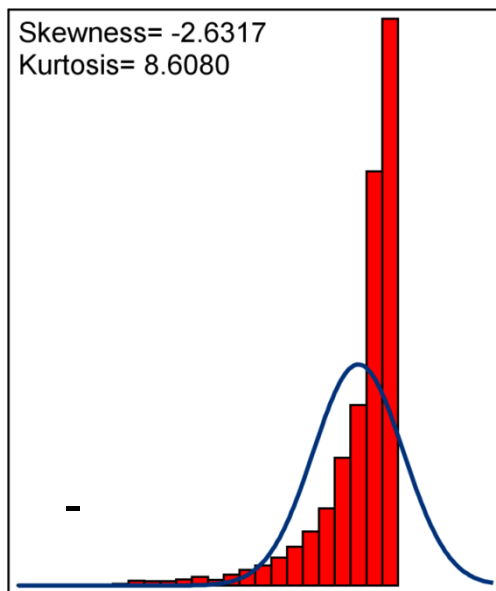
- ◆ A histogram of data from a sample drawn from a normal population generally show values of skewness and kurtosis near zero in SAS output.



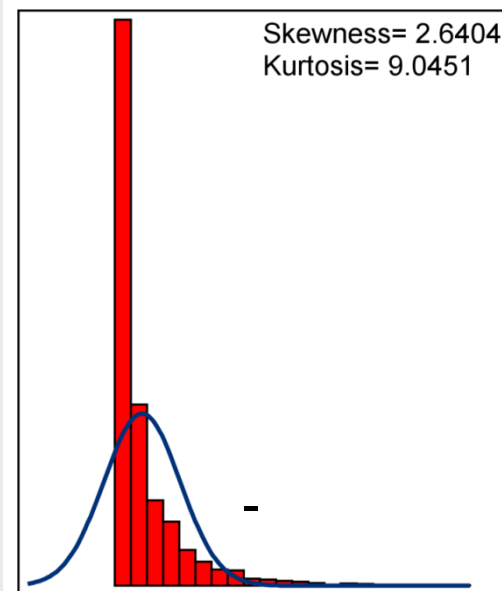
Week 3 Topic:

Skewness

- ◆ One measure of the shape of a distribution is skewness. The *skewness* statistic measures the tendency of your distribution to be more spread out on one side than the other. A distribution that is approximately symmetric has a skewness statistic close to zero.
- ◆ If your distribution is more spread out on the
- ◆ **left** side, then the statistic is negative, and the mean is less than the median. This is sometimes referred to as a *left-skewed* or *negatively skewed* distribution.
- ◆ **right** side, then the statistic is positive, and the mean is greater than the median. This is sometimes referred to as a *right-skewed* or *positively skewed* distribution.



A Left Skewed Distribution

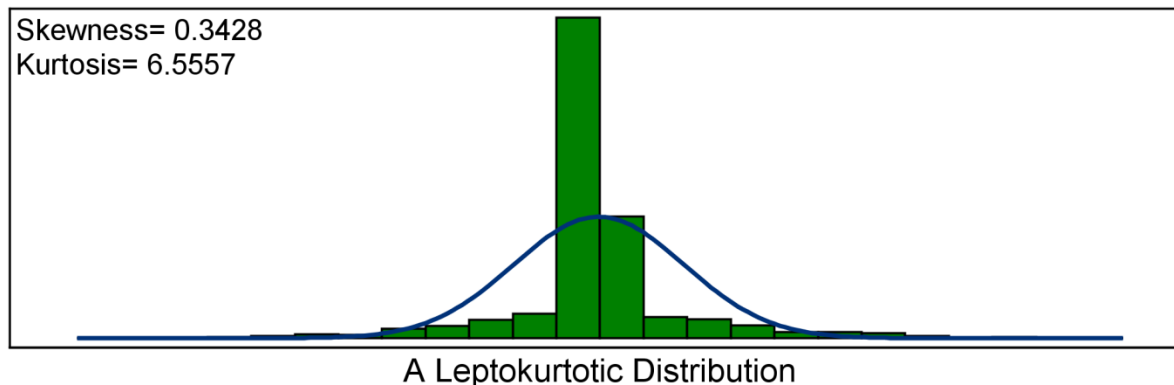
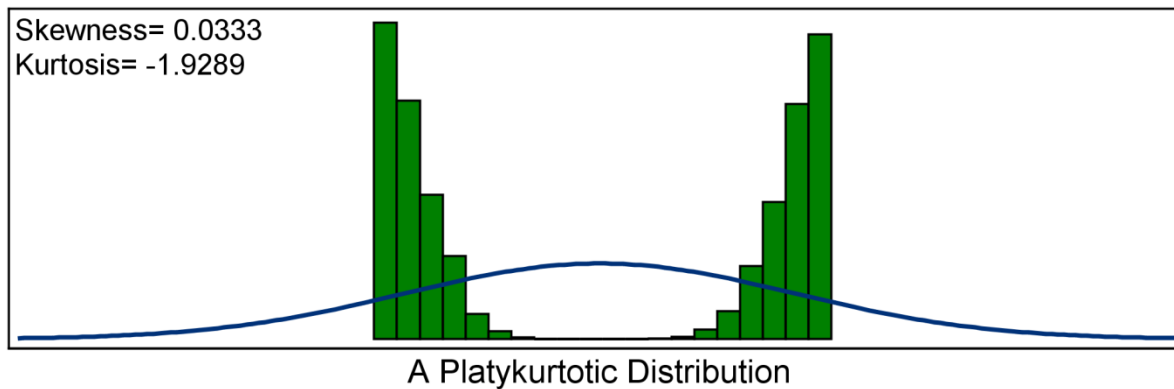


A Right Skewed Distribution

Week 3 Topic:

Kurtosis

- ◆ *Kurtosis* measures the tendency of your data to be distributed toward the center or toward the tails of the distribution. A distribution that is approximately normal has a kurtosis statistic close to zero in SAS. Kurtosis is often very difficult to assess visually.



Week 3 Topic:

Kurtosis (cont.)

- ◆ If the value of your kurtosis statistic is negative, the distribution is said to be *platykurtic*. If the distribution is both symmetric and platykurtic, then there tends to be a smaller-than-normal proportion of observations in the tails and/or a somewhat flat peak. Rectangular, bimodal, and multimodal distributions tend to have low (negative) values of kurtosis.
- ◆ If the value of the kurtosis statistic is positive, the distribution is said to be *leptokurtic*. If the distribution is both symmetric and leptokurtic, then there tends to be a larger-than-normal proportion of observations in the extreme tails and/or a taller peak than the normal. A leptokurtic distribution is often referred to as *heavy-tailed*. Leptokurtic distributions are also sometimes referred to as *outlier-prone distributions*.
- ◆ Distributions that are asymmetric also tend to have nonzero kurtosis. In these cases, understanding kurtosis is considerably more complex than in situations where the distribution is approximately symmetric.

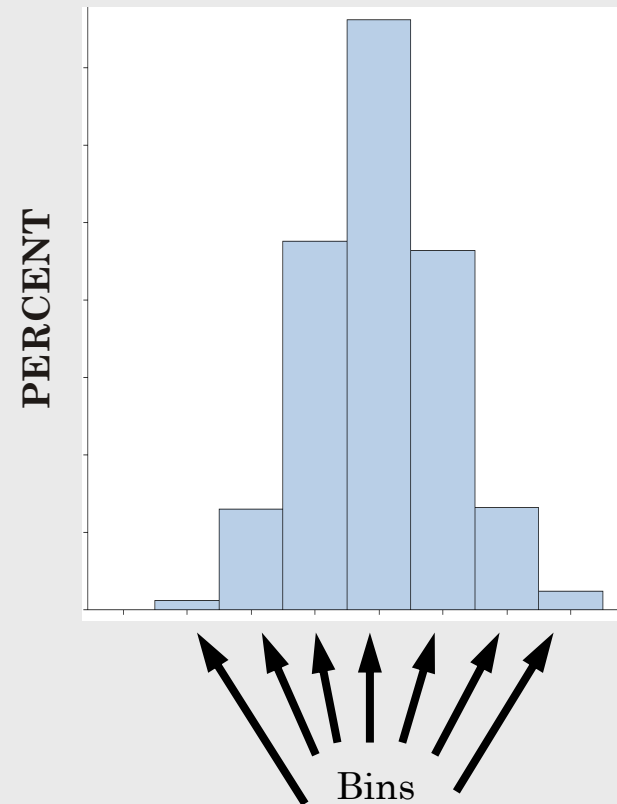
Week 3 Topic:

Graphical Displays of Distributions

- ◆ You can produce the following three types of plots for examining the distribution of your data values:
 - histograms
 - normal probability plots
 - box plots
 - scatter plots

Week 3 Topic: Histograms

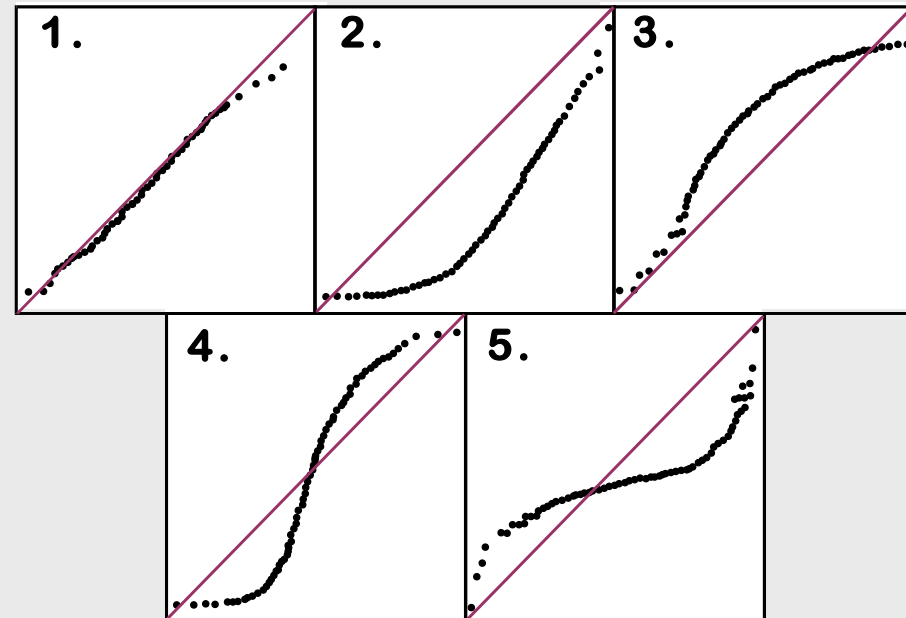
- ◆ Most elementary statistical procedures assume some underlying population probability distribution. It is a good idea to look at your data to see whether the distribution of your sample data can reasonably be assumed to come from a population with the assumed distribution.
- ◆ A histogram is a good way to determine how the probability distribution is shaped.



Week 3 Topic:

Normal Probability Plots

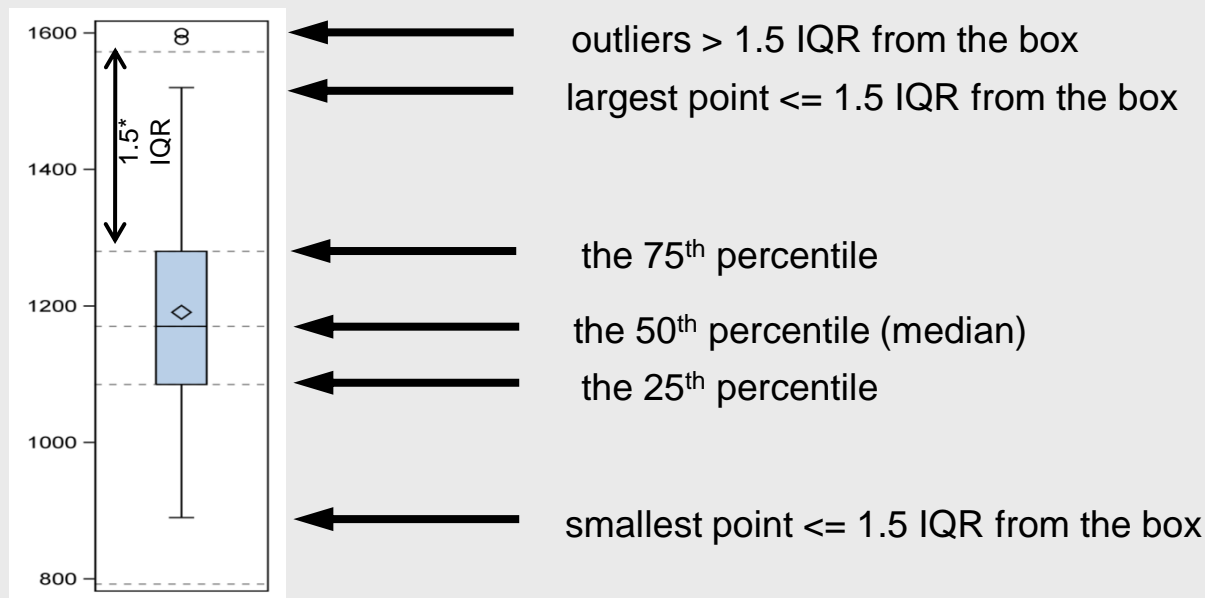
- ◆ A *normal probability plot* is a visual method for determining whether your data comes from a distribution that is approximately normal. The vertical axis represents the actual data values, and the horizontal axis displays the expected percentiles from a standard normal distribution.
- ◆ The above diagrams illustrate some possible normal probability plots for data from the following:
- ◆ normal distribution (The observed data follow the reference line.)
- ◆ skewed-to-the-right distribution
- ◆ skewed-to-the-left distribution
- ◆ light-tailed distribution
- ◆ heavy-tailed distribution



Week 3 Topic:

Box Plots

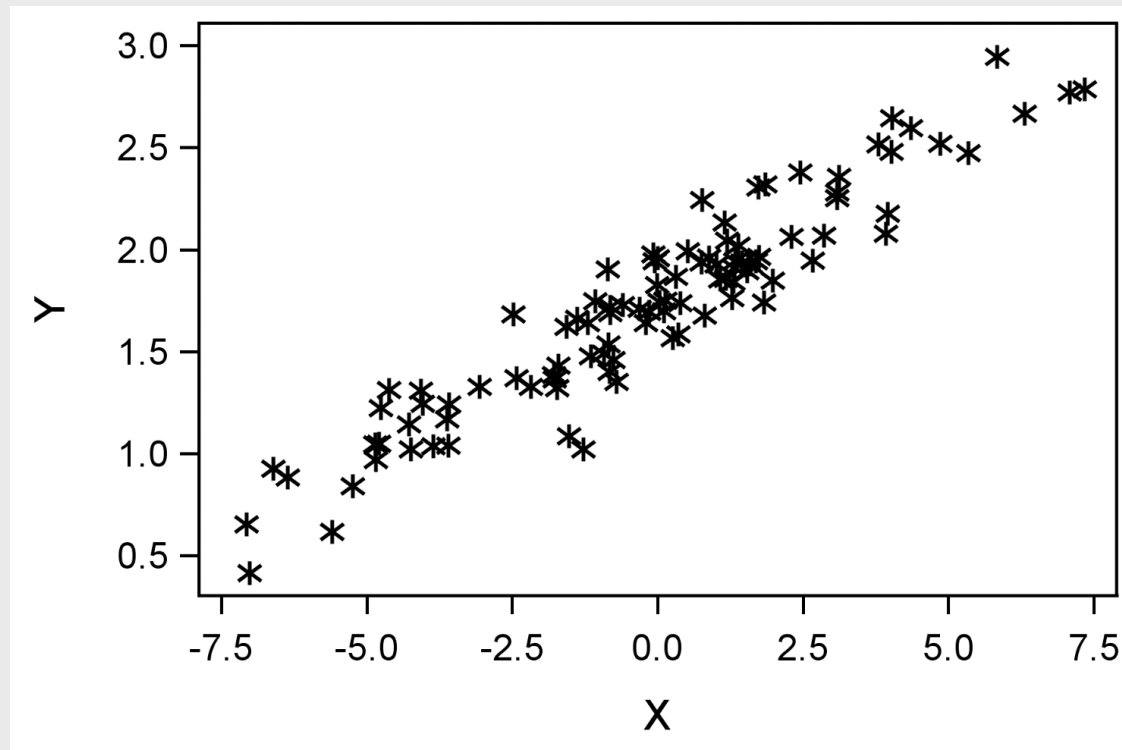
- ◆ *Box plots* (Tukey 1977) (sometimes referred to as *box-and-whisker plots*) provide information about the variability of data and the extreme data values. The box represents the middle 50% of your data (between the 25th and 75th percentile values). You get a rough impression of the symmetry of your distribution by comparing the mean and median, as well as by assessing the symmetry of the box and whiskers around the median line. The whiskers extend from the box as far as the data extends, to a distance of, at most, 1.5 interquartile range (IQR) units. If any values lay more than 1.5 IQR units from either end of the box, they are represented in SAS by individual plot symbols.
- ◆ The plot above shows that the data are approximately symmetric.



The mean is denoted by a \diamond .

Week 3 Topic: Scatter Plots

- ◆ *Scatter plots* are two-dimensional graphs produced by plotting one variable against another within a set of coordinate axes. The coordinates of each point correspond to the values of the two variables.



Week 3 Topic: Scatter Plots (cont.)

- ◆ Scatter plots are useful to accomplish the following:
 - explore the relationships between two variables
 - locate outlying or unusual values
 - identify possible trends
 - identify a basic range of Y and X values
 - communicate data analysis results
- ◆ The predicted value can be thought of as the best estimate of the value of the response at a given value of the predictor variable. Scatter plots show graphically the relationship between predictor variables and response variables.
- ◆ Traditionally, predictor variables are plotted on the x axis and response variables are plotted on the y-axis. A preliminary analysis of associations involves discovery of the presence of associations and their nature.

Week 3 Topic:

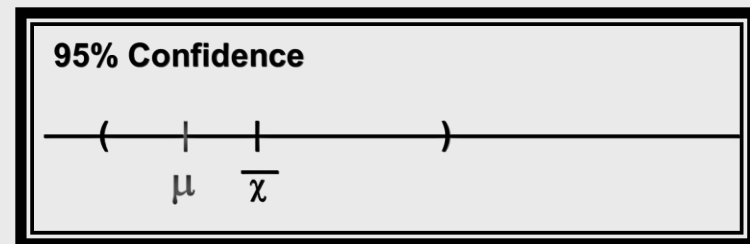
Standard Error of a Mean

- ◆ A statistic that measures the variability of your estimate is the *standard error of the mean*. In statistics, assumptions are often made about distributions of parameters. A common one is that the sampling distribution of parameters is normal. This does not necessarily mean that the units of the population are normally distributed. It is often assumed that the parameter itself is normally distributed. Even though most statisticians only take one sample and get one point estimate for the population parameters, it is useful if they can assume normality of the parameter. The variability of a parameter is measured by its standard error.
- ◆ It differs from the sample standard deviation in that:
 - the sample standard deviation is a measure of the variability of data;
 - the standard error of the mean is a measure of the variability of the sample mean.
 - Standard error of the mean = $\frac{s}{\sqrt{n}} = s_{\bar{x}}$
 - where
 - s is the sample standard deviation.
 - n is the sample size.
- ◆ The standard error of the mean is a measure of precision of the parameter estimate. The smaller the standard error, the more precise your estimate.

Week 3 Topic:

Confidence Interval

- ◆ A *confidence interval* is a range of values that you believe is likely to contain the population parameter of interest is defined by an upper and lower bound around a parameter estimate.
- ◆ To construct a confidence interval, a significance level must be chosen.
- ◆ A 95% confidence interval is commonly used to assess the variability of the sample mean. In the Ames housing sales example, you interpret a 95% confidence interval by stating that you are 95% confident that the interval contains the mean sale price for your population of home sales.
- ◆ You want to be as confident as possible, but remember that if you increase the confidence level too much, the width of your interval increases beyond the point where it is informative. For example, a 100% confidence interval would have confidence bounds of negative and positive infinity.
- ◆ A 95% confidence interval represents a range of values within which you are 95% certain that the true population mean exists.
 - One interpretation is that if 100 different samples were drawn from the same population and 100 intervals were calculated, approximately 95 of them would contain the population mean.



Week 2 Topic:

Overview of Optimization Modeling

Optimization is the process of finding the best values of the variables for a particular criterion or the best decisions for a particular measure of performance.

◆ Components of an Optimization Model:

- Objective – mathematical function to minimize or maximize some measure of performance
- Parameters – Numerical inputs for calculations that may correspond to raw data, estimates, forecasts, or predictions
- Constraints – logical conditions or calculations representing real world limits
- Decision Variables – An unknown quantity or variables to be determined via model run

◆ References for Optimization Modeling:

- Data Smart
- www.solver.com
- <http://opensolver.org>
- Optimization Modeling with Spreadsheets (What's in the PDF handed out. Additional content will be discussed/distributed as needed)
- Step-By-Step_Optimization_S.pdf (in Week 3 Readings)

Week 2 Topic:

Basic Structure of Optimization Modeling

- ◆ The basic structure of a typical mathematical optimization problem formulation is shown here:

min|max objective function

subject to constraints

variable bounds

- ◆ The formulation is easier to understand if it is followed by a description of the decision variables, sets, and parameters.

Week 2 Topic:

Linear Programming Problem

- ◆ Each linear constraint can be either an inequality or an equation
- ◆ Bounds can be $\pm\infty$, so x_j can be restricted to be nonnegative ($l_j = 0$ and $u_j = +\infty$) or free ($l_j = -\infty$ and $u_j = +\infty$)
- ◆ Exhibits proportionality (contribution from any given decision variable to the objective grows in proportion to its value), additivity (contribution from one decision is added to contributions of other decisions), and divisibility (fractional decision variable is meaningful).

$$\begin{array}{ll}\min | \max & f_1x_1 + \dots + f_nx_n \\ \text{subject to} & \mathbf{Ax} \{ \leq, =, \geq \} \mathbf{b} \\ & l_j \leq x_j \leq u_j \quad (j = 1, 2, \dots, n)\end{array}$$

Week 2 Topic:

Nonlinear Programming Problem

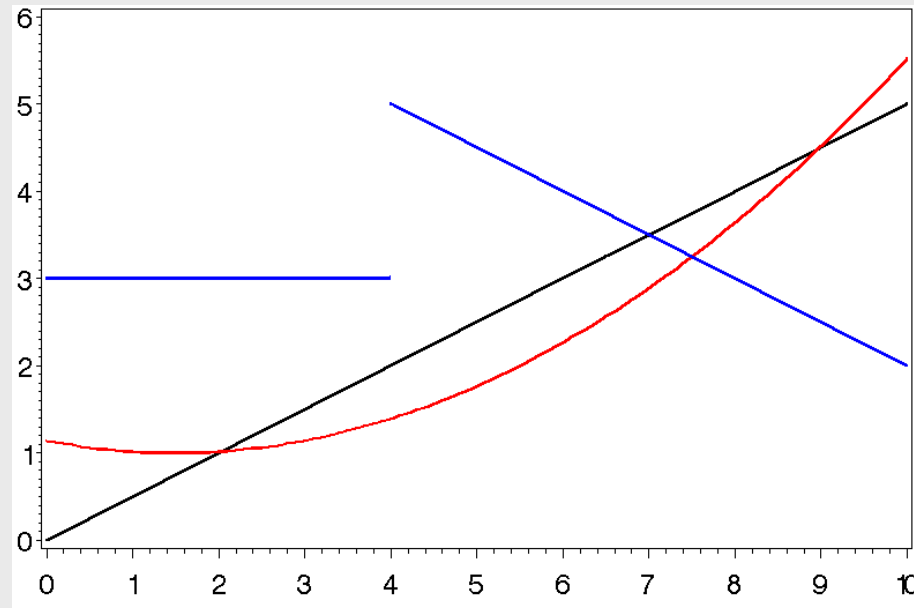
- ◆ $f(\mathbf{x})$ and $c_i(\mathbf{x})$ are continuous functions
- ◆ With the following constraints:
 - **unconstrained:** There are no constraints or bounds.
 - **bound constrained:** There are no constraints other than bounds.
 - **linearly constrained:** All functions $c_i(\mathbf{x})$ are linear.
 - **nonlinearly constrained:** At least one of the functions $c_i(\mathbf{x})$ is nonlinear.

$$\begin{array}{ll}\min | \max & f(\mathbf{x}) \\ \text{subject to} & c_i(\mathbf{x}) \{ \leq, =, \geq \} b_i \quad (i=1,2,\dots,m) \\ & l_j \leq x_j \leq u_j \quad (j=1,2,\dots,n)\end{array}$$

Week 2 Topic:

What's expected using Optimization

- ◆ Able to set up Optimization problems using add in function of Excel
- ◆ Understand the three modeling available:
 - GRG Nonlinear: For nonlinear, smooth (red)
 - Simplex LP: For linear, smooth (black)
 - Evolutionary: For nonlinear, non-smooth (blue)



Week 2 Topic:

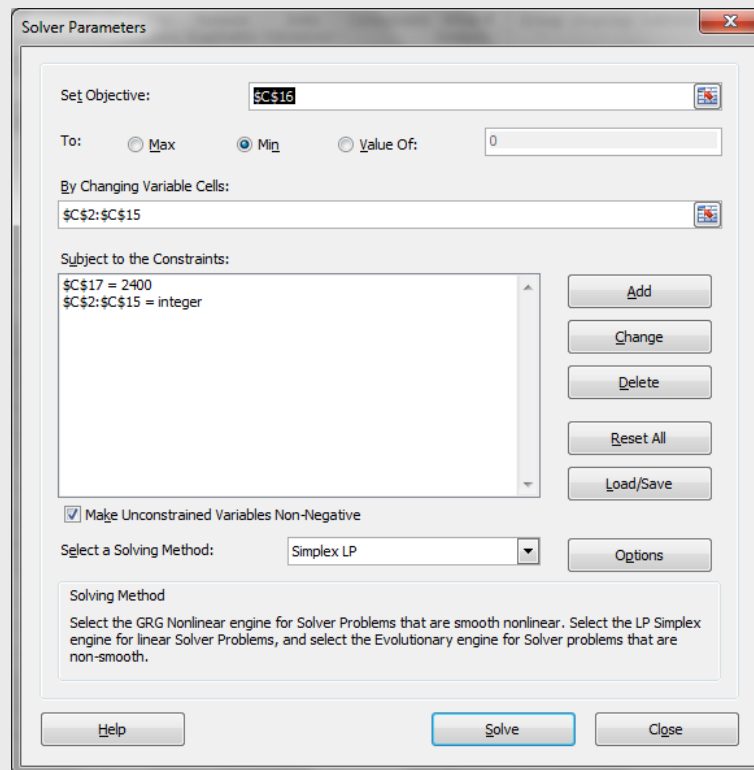
Usage in business

- ◆ **Production Planning:** Determine which of several possible mixes of products should be produced to achieve the highest profit.
- ◆ **Facility Location:** Find the “best” site for, say, a new factory, in relation to the location of materials suppliers, distribution centers, and so on.
- ◆ **Portfolio Selection:** Maximize ROI, balancing return versus risk.
- ◆ **Personnel Assignment:** Match personnel to work requirements in order to meet current needs and anticipated changes, subject to budget and HR requirements.
- ◆ **Supply Chain Planning:** Find the lowest-cost way to move product from factories to distribution centers to stores, and plan for possible disruptions or expansion.
- ◆ **Promotional Marketing:** Determine the best combination of promotional offers, delivery channels, and customers to maximize the overall return on marketing investment.
- ◆ **Supplier Selection and Evaluation:** Choose which suppliers to deal with in order to satisfy requirements and maximize leverage, rating suppliers using a variety of criteria simultaneously.
- ◆ **Inventory Replenishment:** Set inventory policies (reorder levels and maximum stock levels) to meet customer service goals and minimize costs.
- ◆ **Pricing Decisions:** Establish and maintain optimal everyday prices based on costs, regional demand patterns, and competitive price information.

Week 2 Topic:

Back to concession stand example

Using Simplex LP: The model run reflects the book:

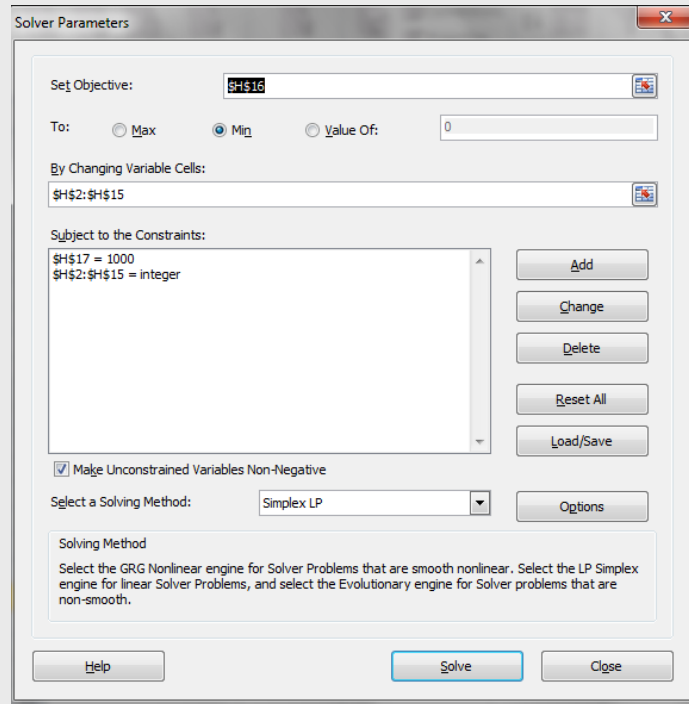


| Item | Calories | How many? |
|-----------------------|----------|-----------|
| Beer | 200 | 0 |
| Bottled Water | 0 | 0 |
| Chocolate Bar | 255 | 0 |
| Chocolate Dipped Cone | 300 | 0 |
| Gummy Bears | 300 | 0 |
| Hamburger | 320 | 0 |
| Hot Dog | 265 | 0 |
| Ice Cream Sandwich | 240 | 0 |
| Licorice Rope | 280 | 1 |
| Nachos | 560 | 2 |
| Pizza | 480 | 0 |
| Popcorn | 500 | 2 |
| Popsicle | 150 | 0 |
| Soda | 120 | 0 |
| Total Items: | | 5 |
| Total Calories: | | 2400 |

Week 2 Topic:

Back to concession stand example

Using Simplex LP: The model says to just sell popcorn.

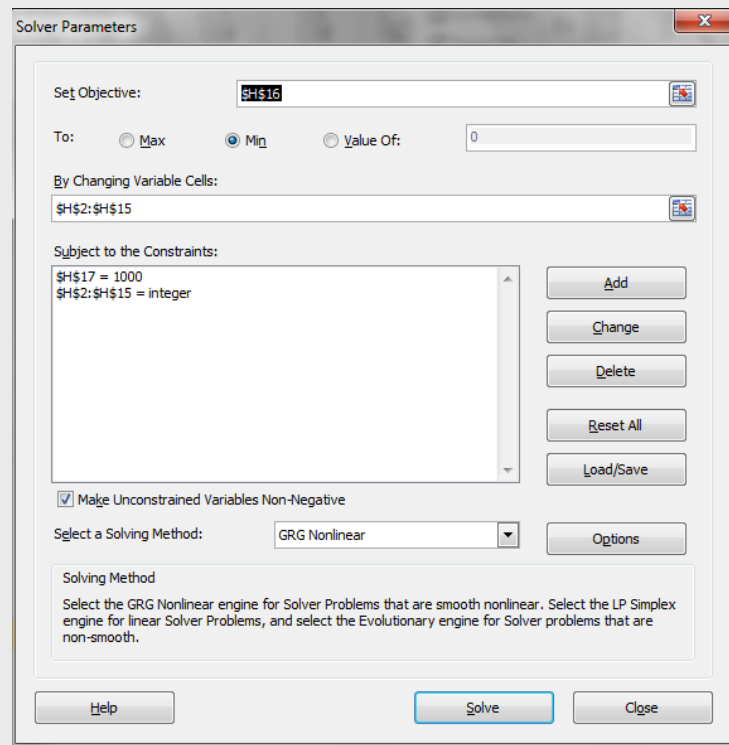


| Item | Unit Price | How many? |
|-----------------------|------------|-----------|
| Popcorn | 5 | 200 |
| Beer | 4 | 0 |
| Bottled Water | 3 | 0 |
| Chocolate Dipped Cone | 3 | 0 |
| Hamburger | 3 | 0 |
| Ice Cream Sandwich | 3 | 0 |
| Nachos | 3 | 0 |
| Popsicle | 3 | 0 |
| Soda | 2.5 | 0 |
| Chocolate Bar | 2 | 0 |
| Gummy Bears | 2 | 0 |
| Licorice Rope | 2 | 0 |
| Pizza | 2 | 0 |
| Hot Dog | 1.5 | 0 |
| Total # of Items: | | 200 |
| Revenue: | | 1000 |

Week 2 Topic:

Back to concession stand example (cont.)

Using GRG Nonlinear: The model says to just sell popcorn. Linear is a subset of nonlinear models. Hence you can solve linear problems using nonlinear models. However, we use linear models on linear problems as it's more dependable.

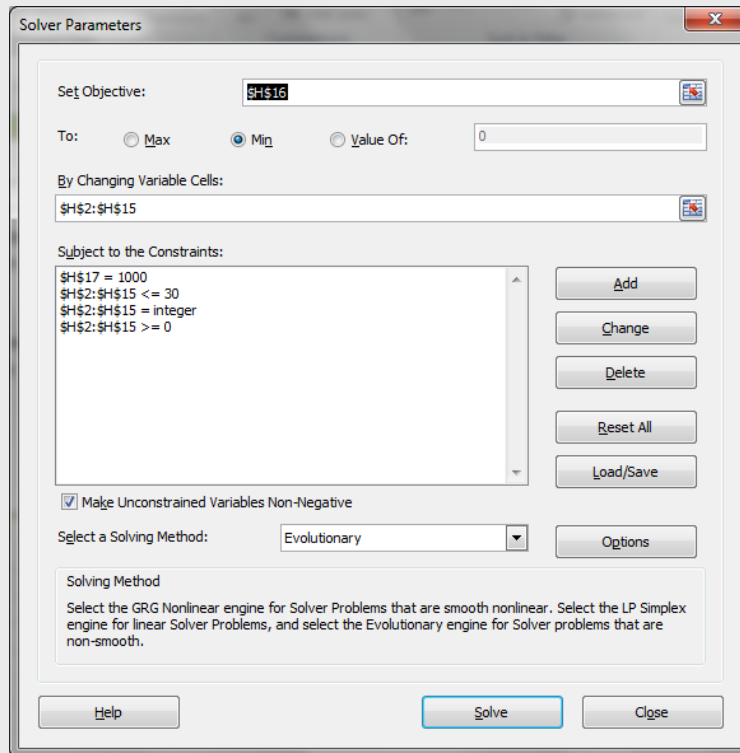


| Item | Unit Price | How many? |
|-----------------------|------------|-----------|
| Popcorn | 5 | 200 |
| Beer | 4 | 0 |
| Bottled Water | 3 | 0 |
| Chocolate Dipped Cone | 3 | 0 |
| Hamburger | 3 | 0 |
| Ice Cream Sandwich | 3 | 0 |
| Nachos | 3 | 0 |
| Popsicle | 3 | 0 |
| Soda | 2.5 | 0 |
| Chocolate Bar | 2 | 0 |
| Gummy Bears | 2 | 0 |
| Licorice Rope | 2 | 0 |
| Pizza | 2 | 0 |
| Hot Dog | 1.5 | 0 |
| Total # of Items: | | 200 |
| Revenue: | | 1000 |

Week 2 Topic:

Back to concession stand example (cont.)

Using Evolutionary and adding bounds to the variables:



| Item | Item Price | Count of item |
|-----------------------|------------|---------------|
| Popcorn | 5 | 30 |
| Beer | 4 | 30 |
| Bottled Water | 3 | 30 |
| Chocolate Dipped Cone | 3 | 30 |
| Hamburger | 3 | 30 |
| Ice Cream Sandwich | 3 | 30 |
| Nachos | 3 | 30 |
| Popsicle | 3 | 30 |
| Soda | 2.5 | 30 |
| Chocolate Bar | 2 | 27 |
| Gummy Bears | 2 | 0 |
| Licorice Rope | 2 | 0 |
| Pizza | 2 | 30 |
| Hot Dog | 1.5 | 0 |
| Total # of items: | | 327 |
| Revenue target: | | 999 |

Week 3 Topic:

Case Study – Pothole Repair Spending

Winter is here. Potholes are a common occurrence in the city of Chicago, especially in the winter. The case study that we will focus on will examine how much the city spends on pothole repairs in the city, what proportion this spending is relative to other city expenses, look for historical trends (4 years of data), and also see if weather conditions have any affect on spending for that year.

We will collect data on:

- ◆ City of Chicago budget appropriations for 2011, 2012, 2013, and 2014
- ◆ Department of transportation's pothole repair spending and performance metrics
- ◆ 311 service request statistics

We will perform analysis to answer questions such as:

- ◆ How much was spent on repairing potholes each year from 2011 to 2014?
- ◆ How much time was spent on repairing potholes?
- ◆ What was the average response times each year?
- ◆ Is the response time seasonally affected e.g., is the response time faster in the winter?

Week 3 Topic:

Case Study – Deliverables

Presentation (PPT):

- 1) Objectives of the analysis
- 2) Data requirements/metadata summary
- 3) Assumptions & Observations on data
- 4) Data validations & Derivations performed
- 6) Analysis details
- 7) Summary of Findings

Data & Analysis (XLS):

- 1) Collection of all raw data (as is)
- 2) Dimensional lookups as necessary (informational as we will not join these)
- 3) Subset of transactions (summary) needed for analysis (apply filters and create worksheet with just the data that you need for the analysis)
- 4) Analysis performed e.g., pivots, charts, graphs, sorts, etc.

Week 3 Topic:

Case Study – Data files to collect

From <http://www.data.gov/>, collect the following data files:

- ◆ Budget - 2011 Budget Ordinance - Positions and Salaries
- ◆ Budget - 2012 Budget Ordinance - Positions and Salaries
- ◆ Budget - 2013 Budget Ordinance - Positions and Salaries
- ◆ Budget - 2014 Budget Ordinance - Positions and Salaries
- ◆ Performance Metrics - Transportation
- ◆ 311 Service Requests - Pot Holes Reported
- ◆ Weather data

Week 3 Topic:

Case Study – Data files to collect

Perform a search on “appropriations Chicago” to get the budget information:

The screenshot shows a web browser window with the URL `catalog.data.gov/dataset?q=appropriations+chicago&sort=score+desc%2C+name+asc&ext_location=&ext_bbox=&ext_prev_extent=-183.515625%2C-56.9`. The browser's address bar and tabs are visible at the top. The Data.gov website interface includes a search bar with the text "Search Data.Gov", a navigation menu with links for DATA, TOPICS, IMPACT, APPLICATIONS, DEVELOPERS, and CONTACT, and a blue header bar with "DATA CATALOG", a home icon, and links for "/ Datasets", "Organizations", and a help icon. The main content area shows a search for "appropriations chicago" with results ordered by Relevance. A filter by location is available, and a map of North America is shown. The search results list 23 datasets found for "appropriations chicago". The first result is "Budget - 2015 Budget Ordinance - Appropriations" by the City of Chicago, described as "The Annual Appropriation Ordinance is the final City operating budget as approved by the City Council. It reflects the City's operating budget at the beginning of...". Below the title are links for CSV, RDF, JSON, and XML. The second result is "Budget - 2012 Budget Recommendations - Appropriations" by the City of Chicago, described as "The dataset details 2012 Budget Recommendations, which is the line-item budget document proposed by the Mayor to the City Council for approval. Budgeted...".

Week 3 Topic:

Case Study – Data files to collect

Perform a search on “potholes Chicago” to get the performance metrics on potholes:

The screenshot shows a web browser window with the URL `catalog.data.gov/dataset?q=potholes+chicago&sort=score+desc%2C+name+asc&ext_location=&ext_bbox=&ext_prev_extent=-183.515625%2C-56.944974`. The browser tabs include "Almanac | Chicago Weat...", "Search for a Dataset - Dat...", and "Data Access | National Ce...". The browser's bookmark bar shows various links like "Chicago area Hadoo...", "Getting started with...", "Suemee Shin's Portf...", "Big Data for Social...", "Rat Hockey & Open...", "Shmoop Site Map", "Package Ice | McFet...", and "Other bookmarks".

The Data.Gov website header features the "DATA.GOV" logo, navigation links for "DATA", "TOPICS", "IMPACT", "APPLICATIONS", "DEVELOPERS", and "CONTACT", and a "DATA CATALOG" section with links for "/ Datasets", "Organizations", and a help icon. A search bar at the top right contains the text "Search Data.Gov".

The main content area shows a search for "potholes chicago". The results are ordered by "Relevance". A filter by location is available, with a map of North and South America. The search results list 3 datasets found for "potholes chicago":

- Performance Metrics - Transportation - Pothole Repair**
City of Chicago – When moisture seeps into pavement, it expands when it freezes and contracts when it thaws. This flexing of the pavement, combined with the melted water and the...
Available formats: CSV, RDF, JSON, XML
- 311 Service Requests - Pot Holes Reported**
City of Chicago – The Chicago Department of Transportation (CDOT) oversees the patching of potholes on over 4,000 miles of arterial and residential streets in Chicago. CDOT receives...
Available formats: CSV, RDF, JSON, XML

The map at the bottom left shows the location of Chicago, with a note: "Map data CC-BY-SA by OpenStreetMap Tiles by MapQuest".

Week 3 Topic:

Weather data

To add weather data, Weather Underground has historical weather data from the National Weather Service, for example:

http://www.wunderground.com/history/airport/KMDW/2013/9/21/CustomHistory.html?dayend=11&monthend=9&yearend=2014&req_city=&req_state=&req_stationname=&reqdb.zip=&reqdb.magic=&reqdb.wmo=

(scroll to the end for a CSV)

Week 3 Topic:

Case Study – Document Assumptions

During the data inspection, transformation, and integration stage, document all observations and assumptions. Also, make sure to collect and review the metadata information provided for each file.

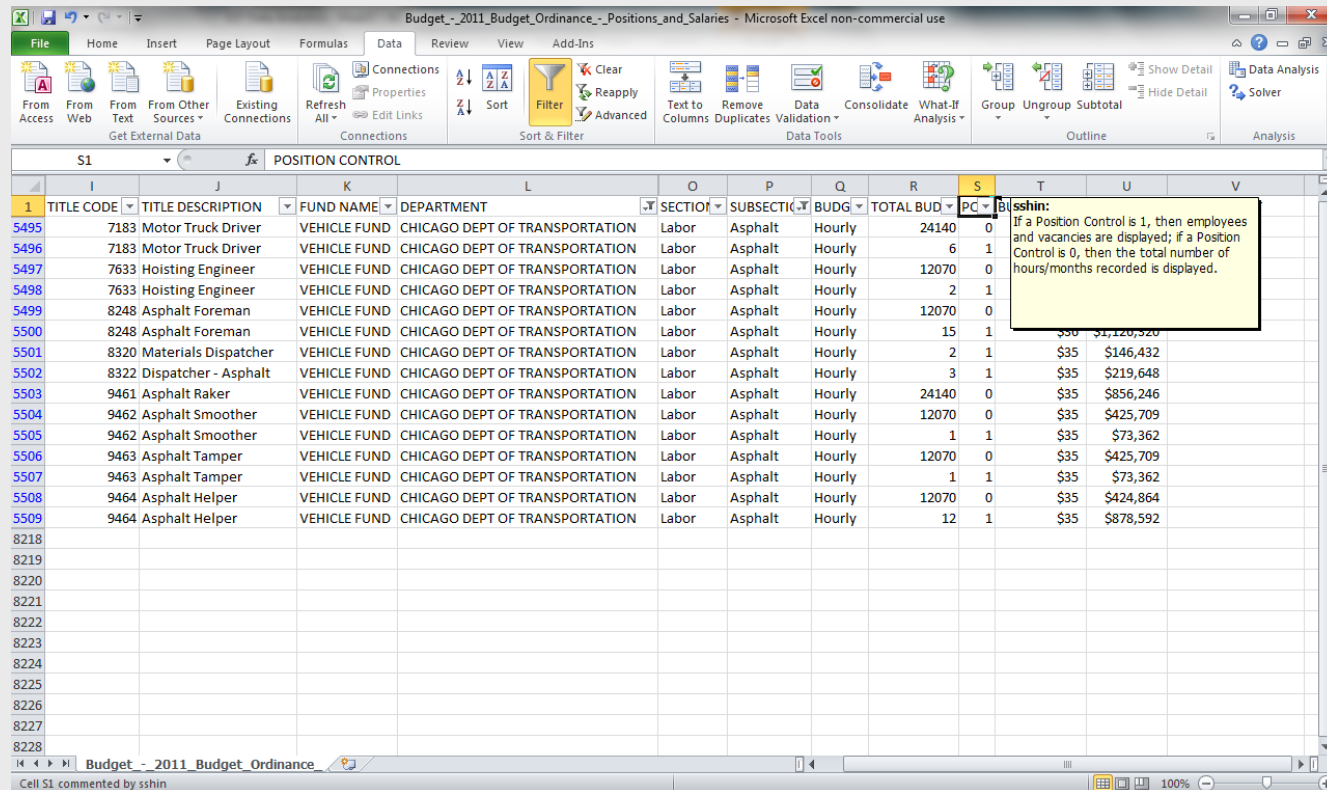
For example, looking at the positions and salaries data for 2011, we make the following assumptions:

- 1) We assume majority of the pothole repair spending is on labor
- 2) Spending for pothole repair is equal to 50% of the spending categorized as “Asphalt” in column SUBSECTION.
- 3) We add POSITION CONTROL ‘0’ and ‘1’ for overall spending. Assumption is that additional contract employee hours are represented as ‘0’

Week 3 Topic:

Case Study – Look at 2011 data

Hid columns A-H, N, and M. Applied filters for columns L, P, and gathered metadata information for POSITION CONTROL:



Budget_-_2011_Budget_Ordinance_-_Positions_and_Salaries - Microsoft Excel non-commercial use

File Home Insert Page Layout Formulas Data Review View Add-Ins

From Access From Web From Text From Other Sources Existing Connections Refresh All Properties Edit Links Connections Sort & Filter Filter Clear Reapply Advanced Text to Columns Remove Duplicates Data Validation Consolidate What-If Analysis Group Ungroup Subtotal Show Detail Hide Detail Data Analysis Solver

S1 POSITION CONTROL

| | I | J | K | L | O | P | Q | R | S | T | U | V |
|------|------------|----------------------|--------------|--------------------------------|---------|------------|--------|-----------|----|----------|-------------|---|
| 1 | TITLE CODE | TITLE DESCRIPTION | FUND NAME | DEPARTMENT | SECTION | SUBSECTION | BUDG | TOTAL BUD | PC | Busslin: | | |
| 5495 | 7183 | Motor Truck Driver | VEHICLE FUND | CHICAGO DEPT OF TRANSPORTATION | Labor | Asphalt | Hourly | 24140 | 0 | | | |
| 5496 | 7183 | Motor Truck Driver | VEHICLE FUND | CHICAGO DEPT OF TRANSPORTATION | Labor | Asphalt | Hourly | 6 | 1 | | | |
| 5497 | 7633 | Hoisting Engineer | VEHICLE FUND | CHICAGO DEPT OF TRANSPORTATION | Labor | Asphalt | Hourly | 12070 | 0 | | | |
| 5498 | 7633 | Hoisting Engineer | VEHICLE FUND | CHICAGO DEPT OF TRANSPORTATION | Labor | Asphalt | Hourly | 2 | 1 | | | |
| 5499 | 8248 | Asphalt Foreman | VEHICLE FUND | CHICAGO DEPT OF TRANSPORTATION | Labor | Asphalt | Hourly | 12070 | 0 | | | |
| 5500 | 8248 | Asphalt Foreman | VEHICLE FUND | CHICAGO DEPT OF TRANSPORTATION | Labor | Asphalt | Hourly | 15 | 1 | \$35 | \$1,120,320 | |
| 5501 | 8320 | Materials Dispatcher | VEHICLE FUND | CHICAGO DEPT OF TRANSPORTATION | Labor | Asphalt | Hourly | 2 | 1 | \$35 | \$146,432 | |
| 5502 | 8322 | Dispatcher - Asphalt | VEHICLE FUND | CHICAGO DEPT OF TRANSPORTATION | Labor | Asphalt | Hourly | 3 | 1 | \$35 | \$219,648 | |
| 5503 | 9461 | Asphalt Raker | VEHICLE FUND | CHICAGO DEPT OF TRANSPORTATION | Labor | Asphalt | Hourly | 24140 | 0 | \$35 | \$856,246 | |
| 5504 | 9462 | Asphalt Smoother | VEHICLE FUND | CHICAGO DEPT OF TRANSPORTATION | Labor | Asphalt | Hourly | 12070 | 0 | \$35 | \$425,709 | |
| 5505 | 9462 | Asphalt Smoother | VEHICLE FUND | CHICAGO DEPT OF TRANSPORTATION | Labor | Asphalt | Hourly | 1 | 1 | \$35 | \$73,362 | |
| 5506 | 9463 | Asphalt Tamber | VEHICLE FUND | CHICAGO DEPT OF TRANSPORTATION | Labor | Asphalt | Hourly | 12070 | 0 | \$35 | \$425,709 | |
| 5507 | 9463 | Asphalt Tamber | VEHICLE FUND | CHICAGO DEPT OF TRANSPORTATION | Labor | Asphalt | Hourly | 1 | 1 | \$35 | \$73,362 | |
| 5508 | 9464 | Asphalt Helper | VEHICLE FUND | CHICAGO DEPT OF TRANSPORTATION | Labor | Asphalt | Hourly | 12070 | 0 | \$35 | \$424,864 | |
| 5509 | 9464 | Asphalt Helper | VEHICLE FUND | CHICAGO DEPT OF TRANSPORTATION | Labor | Asphalt | Hourly | 12 | 1 | \$35 | \$878,592 | |
| 8218 | | | | | | | | | | | | |
| 8219 | | | | | | | | | | | | |
| 8220 | | | | | | | | | | | | |
| 8221 | | | | | | | | | | | | |
| 8222 | | | | | | | | | | | | |
| 8223 | | | | | | | | | | | | |
| 8224 | | | | | | | | | | | | |
| 8225 | | | | | | | | | | | | |
| 8226 | | | | | | | | | | | | |
| 8227 | | | | | | | | | | | | |
| 8228 | | | | | | | | | | | | |

Cell S1 commented by sshin

Week 3 Topic: Case Study – 2011 pivot

Create a pivot to do additional analysis e.g., percentage of Asphalt spending compared to overall budget, resource cost breakout etc.:

The screenshot shows an Excel spreadsheet with a PivotTable summarizing budget data. The PivotTable is structured with 'DEPARTMENT' and 'SUBSECTION NAME' as filters, 'Column Labels' for categories, and 'Row Labels' for specific roles. The values are summed across four columns: TOTAL BUDGETED UNITS, TOTAL BUDGETED AMOUNT, and two additional unlabeled columns.

| Row Labels | Sum of TOTAL BUDGETED UNITS | Sum of TOTAL BUDGETED AMOUNT | Sum of TOTAL BUDGETED UNITS | Sum of TOTAL BUDGETED AMOUNT |
|----------------------|-----------------------------|------------------------------|-----------------------------|------------------------------|
| Asphalt Foreman | 12070 | \$435,727 | 15 | |
| Asphalt Helper | 12070 | \$424,864 | 12 | |
| Asphalt Raker | 24140 | \$856,246 | | |
| Asphalt Smoother | 12070 | \$425,709 | 1 | |
| Asphalt Tamper | 12070 | \$425,709 | 1 | |
| Dispatcher - Asphalt | | | 3 | |
| Hoisting Engineer | 12070 | \$528,666 | 2 | |
| Materials Dispatcher | | | 2 | |
| Motor Truck Driver | 24140 | | 6 | |
| Grand Total | 108630 | | 42 | |

The PivotTable Field List on the right shows the following configuration:

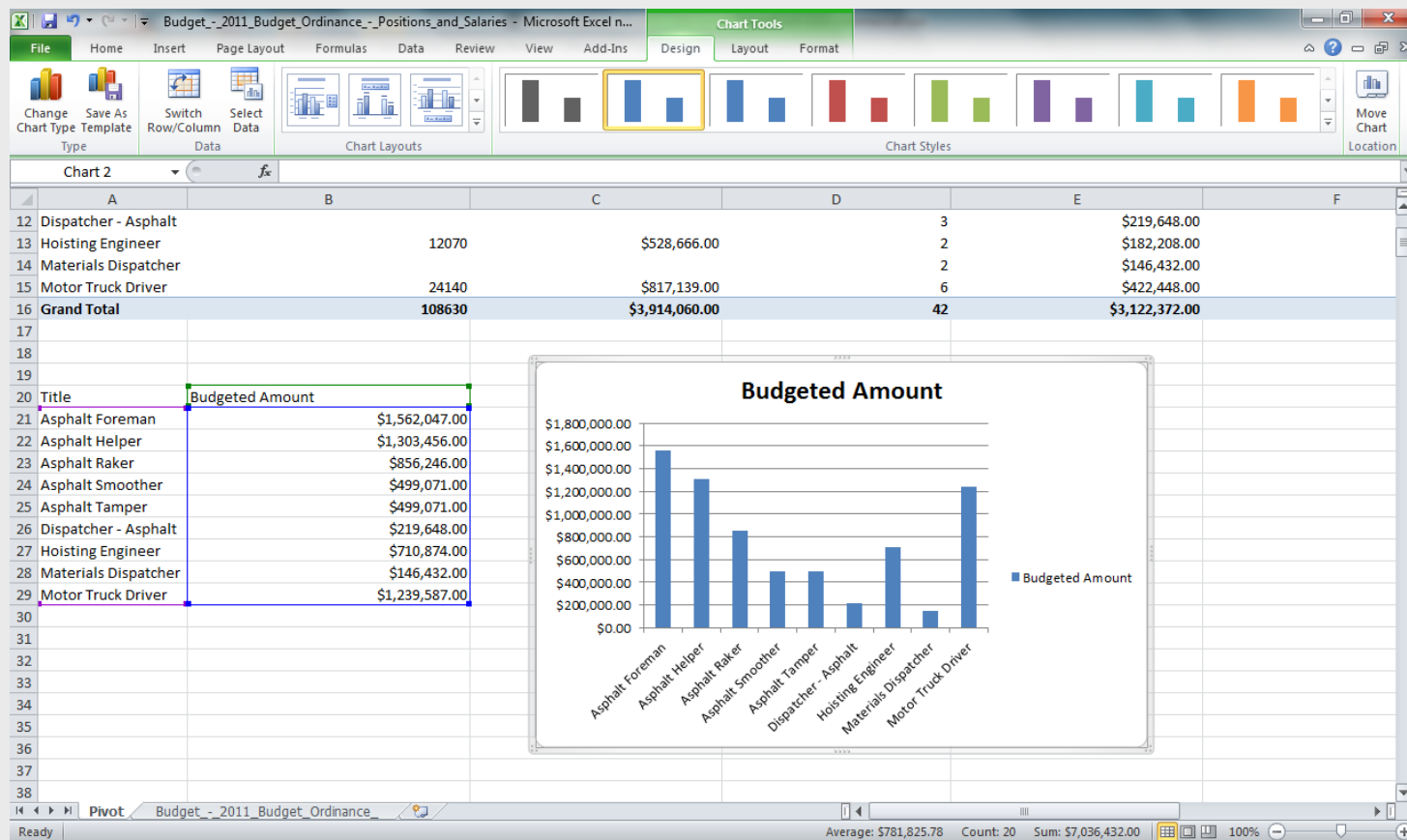
- Report Filter:** DEPARTMENT, SUBSECTION NAME
- Column Labels:** POSITION C..., Σ Values
- Row Labels:** TITLE DESCR..., Sum of TOTA..., Sum of TOTA...

A tooltip for the cell containing \$528,666 shows the calculation: Sum of TOTAL BUDGETED AMOUNT, Value: No value, Row: Materials Dispatcher, Column: 0 - Sum of TOTAL BUDGETED AMOUNT.

ITM - 527

Week 3 Topic: Case Study – 2011 charting

Create charts to understand the spending better:



Week 3 Topic:

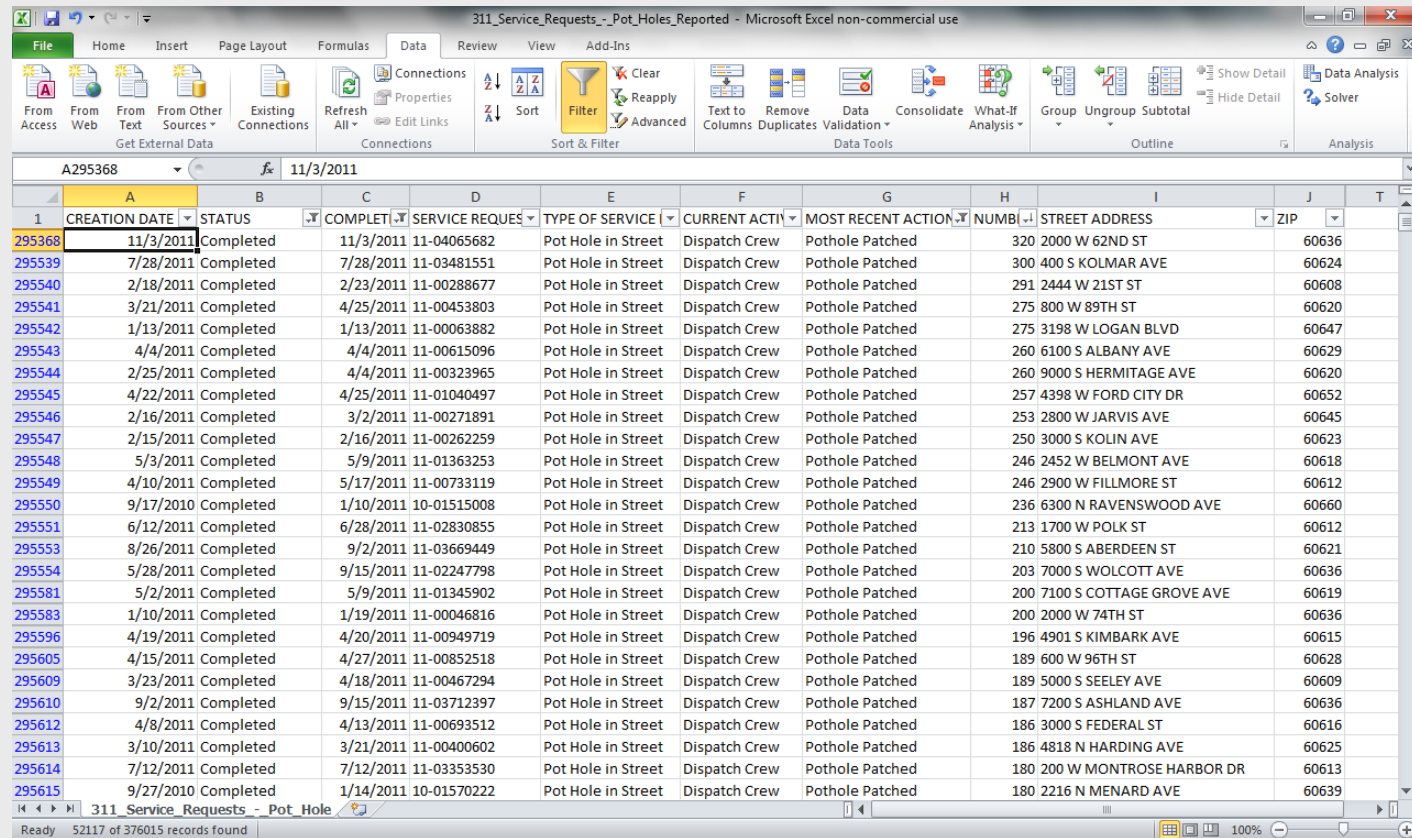
Case Study – 2011 budget metadata

Metadata is in JSON format:

```
{
  "accessLevel": "public",
  "contactPoint": {
    "fn": "<Nobody>",
    "hasEmail": "mailto:",
    "description": "The Annual Appropriation Ordinance is the final City operating budget as approved by the City Council. It reflects the City\u2019s operating budget at the beginning of the fiscal year on January 1, 2011. This dataset displays the positions and related salaries detailed in the budget as of January 1. It is extracted from the personnel portion of the Appropriation Ordinance. The dataset presents the position titles (without names) and salaries described in the budget, but does not provide a reflection of the current city workforce with full names and salaries. Disclaimer: the \u201cTotal Budgeted Units\u201d column displays either A) the number of employees AND vacancies associated with a given position, or B) the number of budgeted units (ie. hours/months) for that position. \u201cPosition Control\u201d determines whether Total Budgeted Units column will count employees and vacancies or hours/months. If a Position Control is 1, then employees and vacancies are displayed; if a Position Control is 0, then the total number of hours/months recorded is displayed. Owner: Budget and Management. Frequency: Data is updated annually. For information on the current city workforce, with names, positions and salaries, visit the \"Current Employee Names, Salaries, and Position Titles\" dataset: http://j.mp/iutKLR.",
    "distribution": [
      {
        "downloadURL": "https://data.cityofchicago.org/api/views/g398-fhbm/rows.csv?accessType=DOWNLOAD",
        "mediaType": "text/csv"
      },
      {
        "downloadURL": "https://data.cityofchicago.org/api/views/g398-fhbm/rows.rdf?accessType=DOWNLOAD",
        "mediaType": "application/rdf+xml"
      },
      {
        "downloadURL": "https://data.cityofchicago.org/api/views/g398-fhbm/rows.json?accessType=DOWNLOAD",
        "mediaType": "application/json"
      },
      {
        "downloadURL": "https://data.cityofchicago.org/api/views/g398-fhbm/rows.xml?accessType=DOWNLOAD",
        "mediaType": "application/xml"
      }
    ],
    "identifier": "https://data.cityofchicago.org/api/views/g398-fhbm",
    "issued": "2011-05-26",
    "keyword": [
      "budget",
      "personnel"
    ],
    "landingPage": "https://data.cityofchicago.org/d/g398-fhbm",
    "modified": "2014-10-28",
    "publisher": {
      "name": "data.cityofchicago.org"
    },
    "theme": [
      "Administration & Finance"
    ],
    "title": "Budget - 2011 Budget Ordinance - Positions and Salaries"
  }
}
```

Week 3 Topic: Case Study – 311 filtered data

Applied filters for columns B, C, and G then sorted by NUMBER OF POTHOLES FILLED ON BLOCK:



| | A | B | C | D | E | F | G | H | I | J | T |
|--------|---------------|-----------|-----------------|--------------------|--------------------|----------------|--------------------|---------------------------|--------------------------|-------|---|
| 1 | CREATION DATE | STATUS | COMPLETION DATE | SERVICE REQUEST ID | TYPE OF SERVICE | CURRENT ACTION | MOST RECENT ACTION | NUMBER OF POTHOLES FILLED | STREET ADDRESS | ZIP | |
| 295368 | 11/3/2011 | Completed | 11/3/2011 | 11-04065682 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 320 | 2000 W 62ND ST | 60636 | |
| 295539 | 7/28/2011 | Completed | 7/28/2011 | 11-03481551 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 300 | 400 S KOLMAR AVE | 60624 | |
| 295540 | 2/18/2011 | Completed | 2/23/2011 | 11-00288677 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 291 | 2444 W 21ST ST | 60608 | |
| 295541 | 3/21/2011 | Completed | 4/25/2011 | 11-00453803 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 275 | 800 W 89TH ST | 60620 | |
| 295542 | 1/13/2011 | Completed | 1/13/2011 | 11-00063882 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 275 | 3198 W LOGAN BLVD | 60647 | |
| 295543 | 4/4/2011 | Completed | 4/4/2011 | 11-00615096 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 260 | 6100 S ALBANY AVE | 60629 | |
| 295544 | 2/25/2011 | Completed | 4/4/2011 | 11-00323965 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 260 | 9000 S HERMITAGE AVE | 60620 | |
| 295545 | 4/22/2011 | Completed | 4/25/2011 | 11-01040497 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 257 | 4398 W FORD CITY DR | 60652 | |
| 295546 | 2/16/2011 | Completed | 3/2/2011 | 11-00271891 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 253 | 2800 W JARVIS AVE | 60645 | |
| 295547 | 2/15/2011 | Completed | 2/16/2011 | 11-00262259 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 250 | 3000 S KOLIN AVE | 60623 | |
| 295548 | 5/3/2011 | Completed | 5/9/2011 | 11-01363253 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 246 | 2452 W BELMONT AVE | 60618 | |
| 295549 | 4/10/2011 | Completed | 5/17/2011 | 11-00733119 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 246 | 2900 W FILLMORE ST | 60612 | |
| 295550 | 9/17/2010 | Completed | 1/10/2011 | 10-01515008 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 236 | 6300 N RAVENSWOOD AVE | 60660 | |
| 295551 | 6/12/2011 | Completed | 6/28/2011 | 11-02830855 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 213 | 1700 W POLK ST | 60612 | |
| 295553 | 8/26/2011 | Completed | 9/2/2011 | 11-03669449 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 210 | 5800 S ABERDEEN ST | 60621 | |
| 295554 | 5/28/2011 | Completed | 9/15/2011 | 11-02247798 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 203 | 7000 S WOLCOTT AVE | 60636 | |
| 295581 | 5/2/2011 | Completed | 5/9/2011 | 11-01345902 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 200 | 7100 S COTTAGE GROVE AVE | 60619 | |
| 295583 | 1/10/2011 | Completed | 1/19/2011 | 11-00046816 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 200 | 2000 W 74TH ST | 60636 | |
| 295596 | 4/19/2011 | Completed | 4/20/2011 | 11-00949719 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 196 | 4901 S KIMBARK AVE | 60615 | |
| 295605 | 4/15/2011 | Completed | 4/27/2011 | 11-00852518 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 189 | 600 W 96TH ST | 60628 | |
| 295609 | 3/23/2011 | Completed | 4/18/2011 | 11-00467294 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 189 | 5000 S SEELEY AVE | 60609 | |
| 295610 | 9/2/2011 | Completed | 9/15/2011 | 11-03712397 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 187 | 7200 S ASHLAND AVE | 60636 | |
| 295612 | 4/8/2011 | Completed | 4/13/2011 | 11-00693512 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 186 | 3000 S FEDERAL ST | 60616 | |
| 295613 | 3/10/2011 | Completed | 3/21/2011 | 11-00400602 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 186 | 4818 N HARDING AVE | 60625 | |
| 295614 | 7/12/2011 | Completed | 7/12/2011 | 11-03353530 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 180 | 200 W MONTROSE HARBOR DR | 60613 | |
| 295615 | 9/27/2010 | Completed | 1/14/2011 | 10-01570222 | Pot Hole in Street | Dispatch Crew | Pothole Patched | 180 | 2216 N MENARD AVE | 60639 | |

Week 3 Topic:

Case Study – 311 metadata

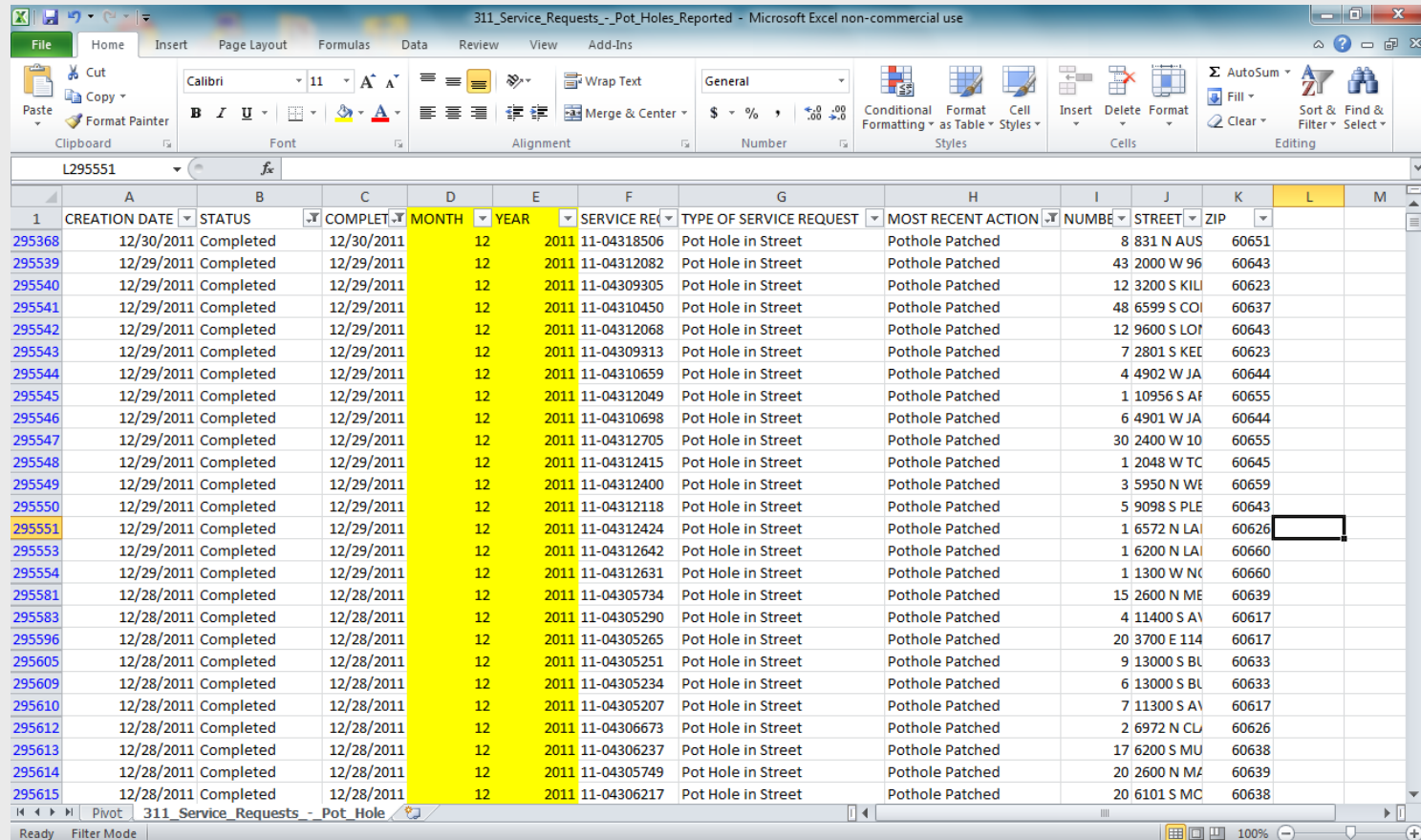
Metadata is in JSON format:

```
{
  "accessLevel": "public",
  "contactPoint": {
    "fn": "<Nobody>",
    "hasEmail": "mailto:",
    "description": "The Chicago Department of Transportation (CDOT) oversees the patching of potholes on over 4,000 miles of arterial and residential streets in Chicago. CDOT receives reports of potholes through the 311 call center and uses a computerized mapping and tracking system to identify pothole locations and efficiently schedule crews. One call to 311 can generate multiple pothole repairs. When a crew arrives to repair a 311 pothole, it fills all the other potholes within the block. Pothole repairs are generally completed within 7 days from the first report of a pothole to 311. Weather conditions, particularly frigid temps and precipitation, influence how long a repair takes. On days when weather is cooperative and there is no precipitation, crews can fill several thousand potholes. \r\n\r\nIf a previous request is already open for a buffer of 4 addresses the request is given the status of \"Duplicate (Open)\". For example, if there is an existing CSR for 6535 N Western and a new request is received for 6531 N Western (which is within four addresses of the original CSR) then the new request is given a status of \"Duplicate (Open)\".\r\n\r\nOnce the street is repaired, the status in CSR will read \u201cCompleted\u201d for the original request and \"Duplicate (Closed)\" for any duplicate requests. A service request also receives the status of \u201cCompleted\u201d when the reported address is inspected but no potholes are found or have already been filled. If another issue is found with the street, such as a \u201ccave-in\u201d or \u201cfailed utility cut\u201d, then it is directed to the appropriate department or contractor. \r\n\r\nData Owner: Transportation. Time Period: All open requests and all completed requests since January 1, 2011. Frequency: Data is updated daily.",
    "distribution": {
      "downloadURL": "https://data.cityofchicago.org/api/views/7as2-ds3y/rows.csv?accessType=DOWNLOAD",
      "mediaType": "text/csv",
      "downloadURL": "https://data.cityofchicago.org/api/views/7as2-ds3y/rows.rdf?accessType=DOWNLOAD",
      "mediaType": "application/rdf+xml",
      "downloadURL": "https://data.cityofchicago.org/api/views/7as2-ds3y/rows.json?accessType=DOWNLOAD",
      "mediaType": "application/json",
      "downloadURL": "https://data.cityofchicago.org/api/views/7as2-ds3y/rows.xml?accessType=DOWNLOAD",
      "mediaType": "application/xml"
    },
    "identifier": "https://data.cityofchicago.org/api/views/7as2-ds3y",
    "issued": "2015-03-25",
    "keyword": ["streets", "pot holes"],
    "landingPage": "https://data.cityofchicago.org/d/7as2-ds3y",
    "modified": "2015-03-25",
    "publisher": {
      "name": "data.cityofchicago.org"
    },
    "theme": ["Service Requests"],
    "title": "311 Service Requests - Pot Holes Reported"
  }
}
```


Week 3 Topic:

Case Study – 311 derived data

To view NUMBER OF POTHOLES FILLED ON BLOCK column data by month, create a month and year column using the MONTH and YEAR functions. Remember to format the new column to *number without decimals* so that the values are displayed correctly:

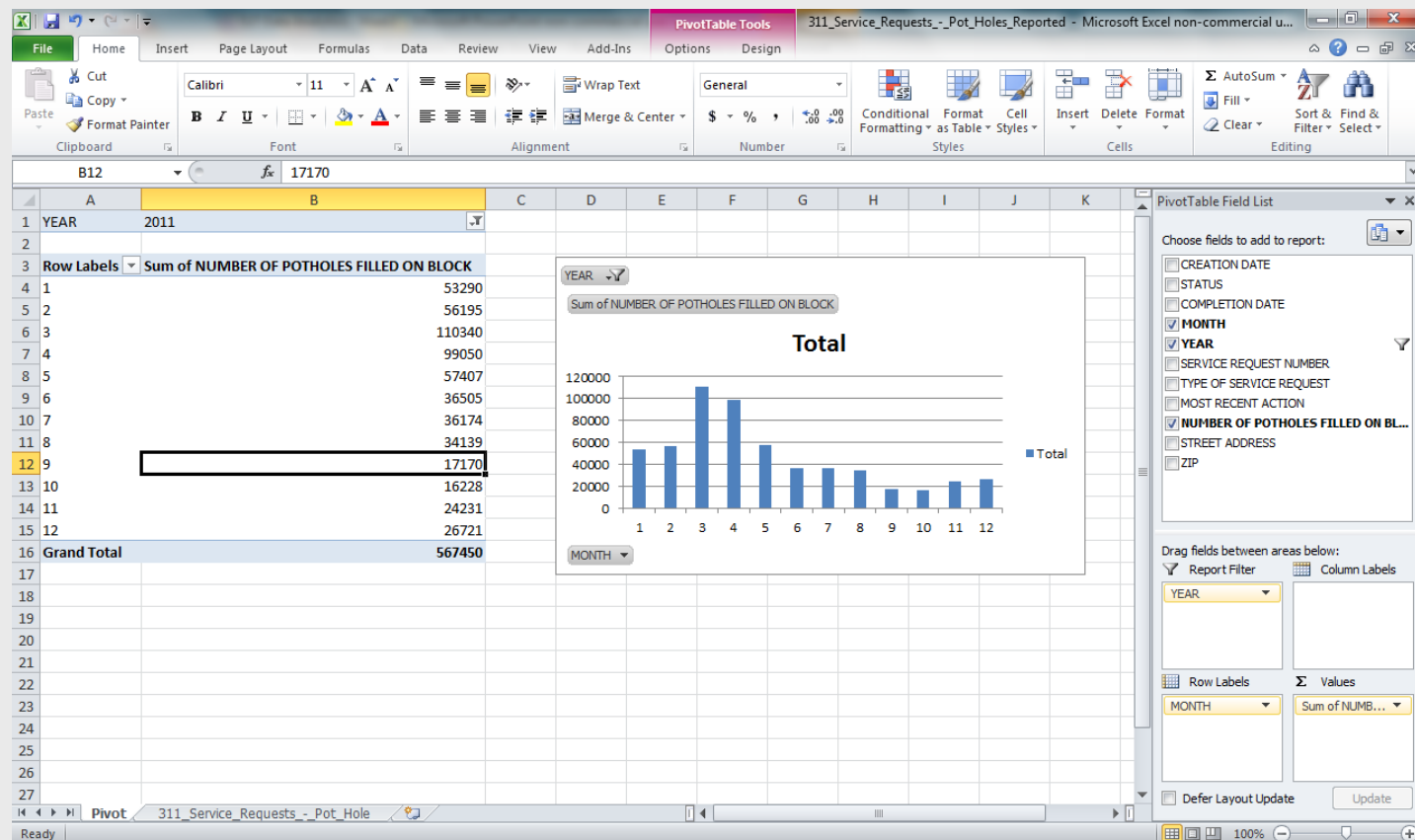


| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|--------|---------------|-----------|------------|-------|------|--------------------|-------------------------|--------------------|------------------------------------|------------|-------|---|---|
| 1 | CREATION DATE | STATUS | COMPLETED | MONTH | YEAR | SERVICE REQUEST ID | TYPE OF SERVICE REQUEST | MOST RECENT ACTION | NUMBER OF POTHOLES FILLED ON BLOCK | STREET | ZIP | | |
| 295368 | 12/30/2011 | Completed | 12/30/2011 | 12 | 2011 | 11-04318506 | Pot Hole in Street | Pothole Patched | 8 | 831 N AUS | 60651 | | |
| 295539 | 12/29/2011 | Completed | 12/29/2011 | 12 | 2011 | 11-04312082 | Pot Hole in Street | Pothole Patched | 43 | 2000 W 96 | 60643 | | |
| 295540 | 12/29/2011 | Completed | 12/29/2011 | 12 | 2011 | 11-04309305 | Pot Hole in Street | Pothole Patched | 12 | 3200 S KIL | 60623 | | |
| 295541 | 12/29/2011 | Completed | 12/29/2011 | 12 | 2011 | 11-04310450 | Pot Hole in Street | Pothole Patched | 48 | 6599 S COI | 60637 | | |
| 295542 | 12/29/2011 | Completed | 12/29/2011 | 12 | 2011 | 11-04312068 | Pot Hole in Street | Pothole Patched | 12 | 9600 S LOI | 60643 | | |
| 295543 | 12/29/2011 | Completed | 12/29/2011 | 12 | 2011 | 11-04309313 | Pot Hole in Street | Pothole Patched | 7 | 2801 S KEE | 60623 | | |
| 295544 | 12/29/2011 | Completed | 12/29/2011 | 12 | 2011 | 11-04310659 | Pot Hole in Street | Pothole Patched | 4 | 4902 W JA | 60644 | | |
| 295545 | 12/29/2011 | Completed | 12/29/2011 | 12 | 2011 | 11-04312049 | Pot Hole in Street | Pothole Patched | 1 | 10956 S AF | 60655 | | |
| 295546 | 12/29/2011 | Completed | 12/29/2011 | 12 | 2011 | 11-04310698 | Pot Hole in Street | Pothole Patched | 6 | 4901 W JA | 60644 | | |
| 295547 | 12/29/2011 | Completed | 12/29/2011 | 12 | 2011 | 11-04312705 | Pot Hole in Street | Pothole Patched | 30 | 2400 W 10 | 60655 | | |
| 295548 | 12/29/2011 | Completed | 12/29/2011 | 12 | 2011 | 11-04312415 | Pot Hole in Street | Pothole Patched | 1 | 2048 W TC | 60645 | | |
| 295549 | 12/29/2011 | Completed | 12/29/2011 | 12 | 2011 | 11-04312400 | Pot Hole in Street | Pothole Patched | 3 | 5950 N WE | 60659 | | |
| 295550 | 12/29/2011 | Completed | 12/29/2011 | 12 | 2011 | 11-04312118 | Pot Hole in Street | Pothole Patched | 5 | 9098 S PLE | 60643 | | |
| 295551 | 12/29/2011 | Completed | 12/29/2011 | 12 | 2011 | 11-04312424 | Pot Hole in Street | Pothole Patched | 1 | 6572 N LAI | 60626 | | |
| 295553 | 12/29/2011 | Completed | 12/29/2011 | 12 | 2011 | 11-04312642 | Pot Hole in Street | Pothole Patched | 1 | 6200 N LAI | 60660 | | |
| 295554 | 12/29/2011 | Completed | 12/29/2011 | 12 | 2011 | 11-04312631 | Pot Hole in Street | Pothole Patched | 1 | 1300 W NC | 60660 | | |
| 295581 | 12/28/2011 | Completed | 12/28/2011 | 12 | 2011 | 11-04305734 | Pot Hole in Street | Pothole Patched | 15 | 2600 N ME | 60639 | | |
| 295583 | 12/28/2011 | Completed | 12/28/2011 | 12 | 2011 | 11-04305290 | Pot Hole in Street | Pothole Patched | 4 | 11400 S AV | 60617 | | |
| 295596 | 12/28/2011 | Completed | 12/28/2011 | 12 | 2011 | 11-04305265 | Pot Hole in Street | Pothole Patched | 20 | 3700 E 114 | 60617 | | |
| 295605 | 12/28/2011 | Completed | 12/28/2011 | 12 | 2011 | 11-04305251 | Pot Hole in Street | Pothole Patched | 9 | 13000 S BL | 60633 | | |
| 295609 | 12/28/2011 | Completed | 12/28/2011 | 12 | 2011 | 11-04305234 | Pot Hole in Street | Pothole Patched | 6 | 13000 S BL | 60633 | | |
| 295610 | 12/28/2011 | Completed | 12/28/2011 | 12 | 2011 | 11-04305207 | Pot Hole in Street | Pothole Patched | 7 | 11300 S AV | 60617 | | |
| 295612 | 12/28/2011 | Completed | 12/28/2011 | 12 | 2011 | 11-04306673 | Pot Hole in Street | Pothole Patched | 2 | 6972 N CL | 60626 | | |
| 295613 | 12/28/2011 | Completed | 12/28/2011 | 12 | 2011 | 11-04306637 | Pot Hole in Street | Pothole Patched | 17 | 6200 S MU | 60638 | | |
| 295614 | 12/28/2011 | Completed | 12/28/2011 | 12 | 2011 | 11-04305749 | Pot Hole in Street | Pothole Patched | 20 | 2600 N MA | 60639 | | |
| 295615 | 12/28/2011 | Completed | 12/28/2011 | 12 | 2011 | 11-04306217 | Pot Hole in Street | Pothole Patched | 20 | 6101 S MC | 60638 | | |

Week 3 Topic:

Case Study – 311 pivot and charting

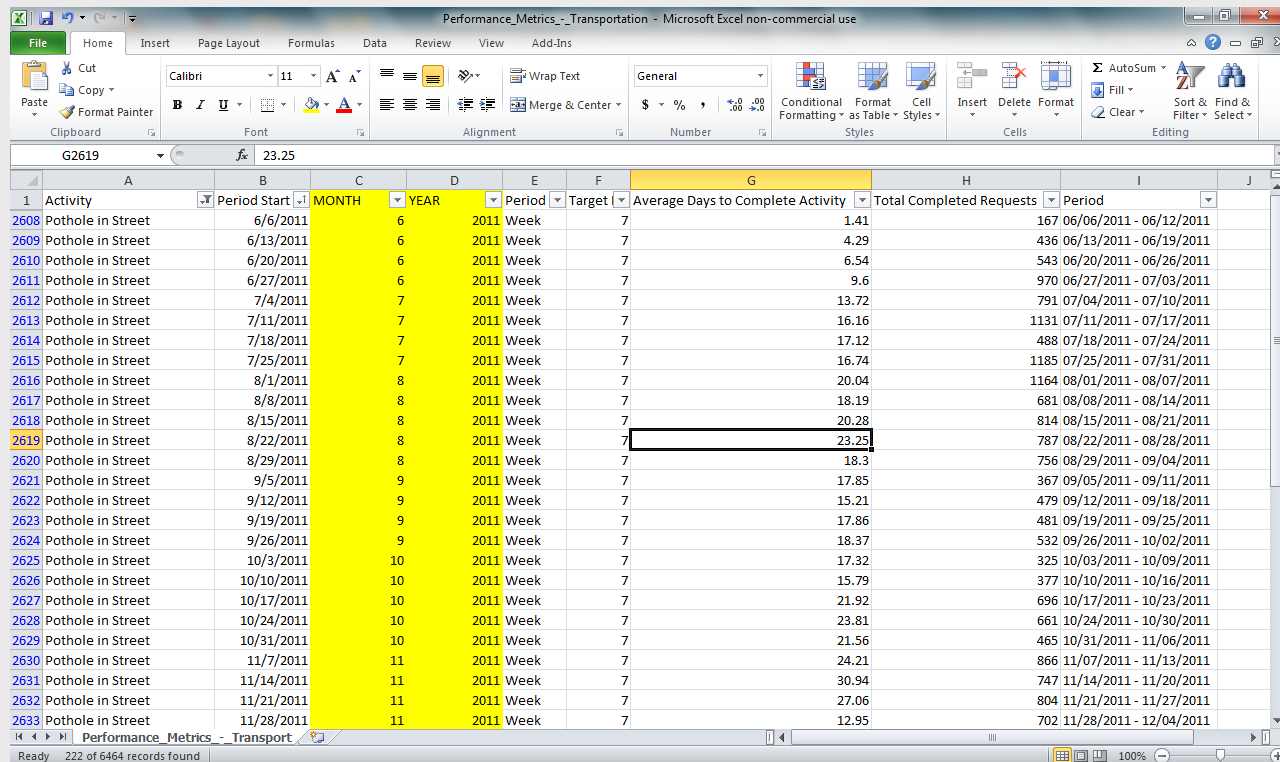
Then, create the following pivot table:



Week 3 Topic:

Case Study – pothole repair metric data

Applied filters for column A and created MONTH and YEAR columns:



| | A | B | C | D | E | F | G | H | I | J |
|------|-------------------|--------------|-------|------|--------|--------|-----------------------------------|--------------------------|-------------------------|---|
| 1 | Activity | Period Start | MONTH | YEAR | Period | Target | Average Days to Complete Activity | Total Completed Requests | Period | |
| 2608 | Pothole in Street | 6/6/2011 | 6 | 2011 | Week | 7 | 1.41 | 167 | 06/06/2011 - 06/12/2011 | |
| 2609 | Pothole in Street | 6/13/2011 | 6 | 2011 | Week | 7 | 4.29 | 436 | 06/13/2011 - 06/19/2011 | |
| 2610 | Pothole in Street | 6/20/2011 | 6 | 2011 | Week | 7 | 6.54 | 543 | 06/20/2011 - 06/26/2011 | |
| 2611 | Pothole in Street | 6/27/2011 | 6 | 2011 | Week | 7 | 9.6 | 970 | 06/27/2011 - 07/03/2011 | |
| 2612 | Pothole in Street | 7/4/2011 | 7 | 2011 | Week | 7 | 13.72 | 791 | 07/04/2011 - 07/10/2011 | |
| 2613 | Pothole in Street | 7/11/2011 | 7 | 2011 | Week | 7 | 16.16 | 1131 | 07/11/2011 - 07/17/2011 | |
| 2614 | Pothole in Street | 7/18/2011 | 7 | 2011 | Week | 7 | 17.12 | 488 | 07/18/2011 - 07/24/2011 | |
| 2615 | Pothole in Street | 7/25/2011 | 7 | 2011 | Week | 7 | 16.74 | 1185 | 07/25/2011 - 07/31/2011 | |
| 2616 | Pothole in Street | 8/1/2011 | 8 | 2011 | Week | 7 | 20.04 | 1164 | 08/01/2011 - 08/07/2011 | |
| 2617 | Pothole in Street | 8/8/2011 | 8 | 2011 | Week | 7 | 18.19 | 681 | 08/08/2011 - 08/14/2011 | |
| 2618 | Pothole in Street | 8/15/2011 | 8 | 2011 | Week | 7 | 20.28 | 814 | 08/15/2011 - 08/21/2011 | |
| 2619 | Pothole in Street | 8/22/2011 | 8 | 2011 | Week | 7 | 23.25 | 787 | 08/22/2011 - 08/28/2011 | |
| 2620 | Pothole in Street | 8/29/2011 | 8 | 2011 | Week | 7 | 18.3 | 756 | 08/29/2011 - 09/04/2011 | |
| 2621 | Pothole in Street | 9/5/2011 | 9 | 2011 | Week | 7 | 17.85 | 367 | 09/05/2011 - 09/11/2011 | |
| 2622 | Pothole in Street | 9/12/2011 | 9 | 2011 | Week | 7 | 15.21 | 479 | 09/12/2011 - 09/18/2011 | |
| 2623 | Pothole in Street | 9/19/2011 | 9 | 2011 | Week | 7 | 17.86 | 481 | 09/19/2011 - 09/25/2011 | |
| 2624 | Pothole in Street | 9/26/2011 | 9 | 2011 | Week | 7 | 18.37 | 532 | 09/26/2011 - 10/02/2011 | |
| 2625 | Pothole in Street | 10/3/2011 | 10 | 2011 | Week | 7 | 17.32 | 325 | 10/03/2011 - 10/09/2011 | |
| 2626 | Pothole in Street | 10/10/2011 | 10 | 2011 | Week | 7 | 15.79 | 377 | 10/10/2011 - 10/16/2011 | |
| 2627 | Pothole in Street | 10/17/2011 | 10 | 2011 | Week | 7 | 21.92 | 696 | 10/17/2011 - 10/23/2011 | |
| 2628 | Pothole in Street | 10/24/2011 | 10 | 2011 | Week | 7 | 23.81 | 661 | 10/24/2011 - 10/30/2011 | |
| 2629 | Pothole in Street | 10/31/2011 | 10 | 2011 | Week | 7 | 21.56 | 465 | 10/31/2011 - 11/06/2011 | |
| 2630 | Pothole in Street | 11/7/2011 | 11 | 2011 | Week | 7 | 24.21 | 866 | 11/07/2011 - 11/13/2011 | |
| 2631 | Pothole in Street | 11/14/2011 | 11 | 2011 | Week | 7 | 30.94 | 747 | 11/14/2011 - 11/20/2011 | |
| 2632 | Pothole in Street | 11/21/2011 | 11 | 2011 | Week | 7 | 27.06 | 804 | 11/21/2011 - 11/27/2011 | |
| 2633 | Pothole in Street | 11/28/2011 | 11 | 2011 | Week | 7 | 12.95 | 702 | 11/28/2011 - 12/04/2011 | |

Week 3 Topic:

Case Study – pothole repair metric pivot

Pivot of transportation metric data. Some data is missing in 2011:

The screenshot shows an Excel spreadsheet with a PivotTable summarizing pothole repair data. The PivotTable is located in the range B3:D11. The PivotTable Field List task pane is open on the right side of the screen.

PivotTable Data:

| Row Labels | Average of Average Days to Complete Activity | Sum of Total Completed Requests |
|--------------------|--|---------------------------------|
| 6 | 5.46 | 2116 |
| 7 | 15.935 | 3595 |
| 8 | 20.012 | 4202 |
| 9 | 17.3225 | 1859 |
| 10 | 20.08 | 2524 |
| 11 | 23.79 | 3119 |
| 12 | 14.425 | 2879 |
| Grand Total | 16.93966667 | 20294 |

PivotTable Field List:

- Choose fields to add to report:
 - ☐ Activity
 - ☐ Period Start
 - ☒ MONTH
 - ☒ YEAR
 - ☐ Period Length
 - ☐ Target Response Days
 - ☒ Average Days to Complete Activity
 - ☒ Total Completed Requests
- Drag fields between areas below:
 - Report Filter: YEAR
 - Column Labels: Values (Sum of Total Completed Requests)
 - Row Labels: MONTH (Average of Average Days to Complete Activity)

Week 3 Topic:

Case Study – transportation metadata

Metadata is in JSON format:

```
{
  "accessLevel": "public",
  "contactPoint": {
    "fn": "<Nobody>",
    "hasEmail": "mailto:dataportal@cityofchicago.org"
  },
  "description": "When moisture seeps into pavement, it expands when it freezes and contracts when it thaws. This flexing of the pavement, combined with the melted water and the stress of vehicular traffic, causes pavement to deteriorate and potholes to form. The Department of Transportation (CDOT) responds to potholes reported through 311's Customer Service Requests (CSR) system by mapping open pothole requests each morning and routing crews in geographic clusters so as to fill as many potholes as possible per day. This metric tracks the average number of days CDOT takes to complete pothole repairs per week. Total number of requests fulfilled per week is also available by mousing over columns. The target response time for pothole repairs is within 7 days. For more information about pothole repairs, see CDOT's pothole Frequently Asked Questions page: http://www.cityofchicago.org/content/dam/city/depts/cdot/PotholeFAQ_winter1011.pdf",
  "distribution": [
    {
      "downloadURL": "https://data.cityofchicago.org/api/views/eaff-5ff2/rows.csv?accessType=DOWNLOAD",
      "mediaType": "text/csv"
    },
    {
      "downloadURL": "https://data.cityofchicago.org/api/views/eaff-5ff2/rows.rdf?accessType=DOWNLOAD",
      "mediaType": "application/rdf+xml"
    },
    {
      "downloadURL": "https://data.cityofchicago.org/api/views/eaff-5ff2/rows.json?accessType=DOWNLOAD",
      "mediaType": "application/json"
    },
    {
      "downloadURL": "https://data.cityofchicago.org/api/views/eaff-5ff2/rows.xml?accessType=DOWNLOAD",
      "mediaType": "application/xml"
    }
  ],
  "identifier": "https://data.cityofchicago.org/api/views/eaff-5ff2",
  "issued": "2015-07-15",
  "keyword": ["service requests", "pothole", "performance metrics", "streets", "pavement"],
  "landingPage": "https://data.cityofchicago.org/d/eaff-5ff2",
  "modified": "2015-09-02",
  "publisher": {
    "name": "data.cityofchicago.org",
    "theme": ["Administration & Finance"]
  },
  "title": "Performance Metrics - Transportation"
}
```

Week 3 Topic:

Data Requirements

A spreadsheet listing of data attribute and information on the attributes including the following as columns:

- ◆ **Subject Area:** Category, grouping of like information e.g., Budget
- ◆ **Table/Entity Name:** e.g., department, position, etc.
- ◆ **Element/Attribute Name:** this can be conceptual/logical name or physical name as depicted in a database table. For our case, we will focus on conceptual/logical names
- ◆ **Description:** a short paragraph or sentence describing the attribute
- ◆ **Data Type:** number, text, currency, etc.
- ◆ **Valid Range:** depicts a continuous numeric range of values e.g., number of resources should be greater than 0 with an upper bound and integer
- ◆ **Value list:** To collect value lists of items from raw data, use pivot tables
- ◆ **Usage:** describes what the data will be used for e.g., performance metric, budget amount, etc. In some cases, this can be downstream applications
- ◆ **History required:** describes how much historical data to store and at what granularity e.g., 2011-2014 years of data by year or month dependent on analysis. Our 311 raw data is transactional hence has dates associated with each report, use pivot table to get month summaries

Week 3 Topic:

Metadata example

| Column Name | Description |
|------------------------------------|--|
| YEAR | Year to which budgeting data belongs |
| FUND CODE | Source system generated unique code for each type of fund |
| FUND NAME | Type of fund |
| DEPARTMENT CODE | Source system generated unique code for each department |
| DEPARTMENT NAME | List of departments like water mgmt, construction, electricity etc. |
| DIVISION CODE | Source system generated unique code for each division |
| DIVISION NAME | Division is one step below the department |
| SECTION CODE | Source system generated unique code for each section |
| SECTION NAME | Section falls under division. It can be admin, finance, labor etc. |
| SUBSECTION CODE | Source system generated unique code for each type of sub-section |
| SUBSECTION NAME | Sub-section falls under section. More granular level. |
| TITLE CODE | Source system generated unique code for each title |
| TITLE DESCRIPTION | This is most granular level in hierarchy which depicts role with respect to section and sub-sections |
| BUDGETED UNIT | Unit - Monthly/yearly/hourly |
| TOTAL BUDGETED UNITS | Count of units |
| POSITION CONTROL | If a Position Control is 1, then employees and vacancies are displayed; if a Position Control is 0, then the total number of hours/months recorded is displayed. |
| BUDGETED PAY RATE | Rate associated with title |
| TOTAL BUDGETED AMOUNT | Total amount spent |
| SERVICE REQUEST NUMBER | Generated number with respect to the request raised by citizen |
| TYPE OF SERVICE REQUEST | Request can be for pothole patching, resurfacing etc. |
| MOST RECENT ACTION | Action taken with respect to each request |
| NUMBER OF POTHoles FILLED ON BLOCK | Number of potholes patched or resurfaced |
| CREATION DATE | Request created on |
| STATUS | Status of each request |
| COMPLETION DATE | Request completed on |
| STREET ADDRESS | Address information where pothole is present |
| ZIP | ZIP code of area |
| Activity | List of activities performed to resolve public facilities' problem like cable cut, pothole patching, traffic problem etc. |
| Period Start | Weekly period of request |
| Target Response Days | Maximum time required to complete request |
| Average Days to Complete Activity | Time taken to complete request |
| Total Completed Requests | Number of requests completed |

Week 3 Topic:

Pothole analysis – PPT Layout

Presentation (PPT):

- 1) Introduction - objectives
- 2) Background – maps, metadata, data characteristics, data issues, etc
- 3) Findings I – spending, performance metrics analysis by year
- 4) Findings II - % of overall spending by year
- 5) Other findings e.g., weather related analysis
- 6) Summary

Week 3 Topic:

PPT Layout - introduction

- ◆ Include objectives. Tell the reader what we are looking at and why we are doing the analysis. For example:

29 states and 7 union territories exist in the country of India. India uses pincodes to manage postal services. We look at pincodes in these states and territories, find which has the most codes for servicing and understand geographic density. Analysis will be performed at the geographic region (postal zone) and state level.

- ◆ Include data source information
- ◆ Highlight information about the data as necessary
- ◆ Highlight statistics on data files, data content, or other relevant metrics on data

Week 3 Topic:

PPT Layout - background

- ◆ List data used versus not. State reasons why
- ◆ Include statistics on data files. For example:

Statistics for the download file:

Total # of rows in file: 154797

Total # of geographic regions (postal code) in dataset: 8

Total # of States in dataset: 36

- ◆ Include relevant facts about the data. For example:

Postal Index Number (PIN) or PIN Code is a 6 digit code of Post Office numbering used by India Post. The PIN was introduced on August 15, 1972. There are 9 PIN regions in the country. The first 8 are geographical regions and the digit 9 is reserved for the Army Postal Service. The first digit indicates one of the regions. The first 2 digits together indicate the sub region or one of the postal circles. The first 3 digits together indicate a sorting / revenue district. The last 3 digits refer to the delivery Post Office.

- ◆ Include any data issues
- ◆ Include a map or pictorial description of the data
- ◆ Gather relevant information outside of the data file to add context to the data file

Week 3 Topic:

PPT Layout - findings

- ◆ Follow the guidelines and include analysis with surrounding text
- ◆ Answer questions on data. For example:

How much was spent on repairing potholes each year from 2011 to 2014?

How much time was spent on repairing potholes?

What was the average response times each year?

Is the response time seasonally affected e.g., is the response time faster in the winter?

- ◆ Highlight relevant assumptions as needed to add context to analysis
- ◆ Highlight derived data as necessary to complete the analysis
- ◆ Add commentary on why certain analysis needs additional data – if that is the case

Week 3 Topic:

PPT Layout - summary

- ◆ Highlight key findings in bullet points.
- ◆ This can be restatements of key points in the Findings slide or additional summary analysis
- ◆ State any next steps in the analysis

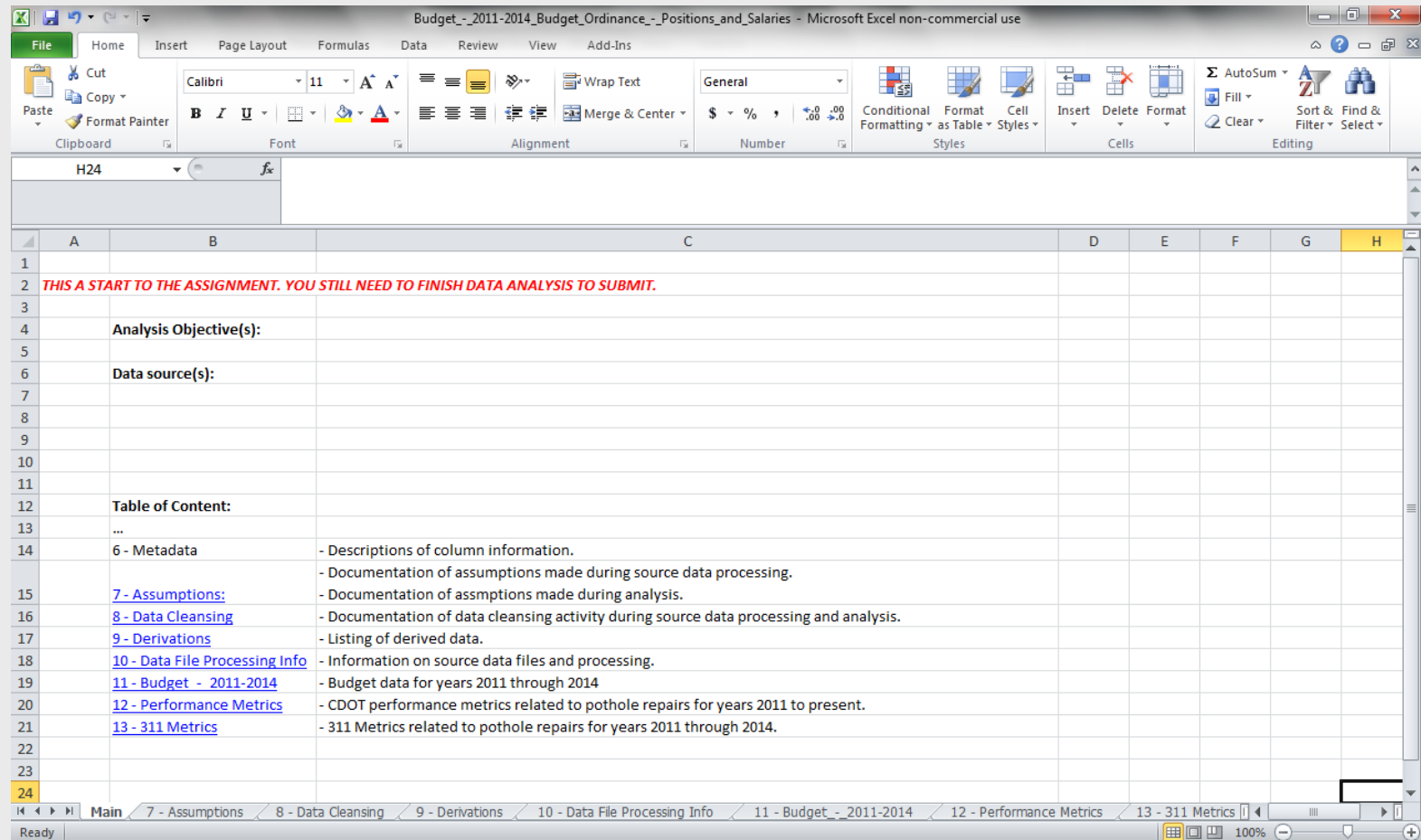
Week 3 Topic:

Pothole analysis – Excel Layout

Data & Analysis (XLS):

- 1) Summary – intro, objectives, TOC, version(s), etc.
- 2) Finding(s) analysis worksheet
- 3) Document metadata
- 4) Document assumptions with links to relevant columns
- 5) List data cleansing/validations performed as necessary with links to relevant columns
- 6) List any derived data with links to relevant columns
- 7) Document data processing
- 8) Data tab(s) for analysis

Week 3 Topic: Example Excel Layout



Review from Week 1: Data Analysis Methodology

- A. Inspect - All data is inspected and “cleansed”
 - Records are investigated and fixed where appropriate e.g., outliers, type check, range check
 - Consistent default values are assigned to “missing” data
 - Data validation codes are included in the data where appropriate (e.g. invalid zip code) to avoid/compensate/exclude during analysis
- B. Transform - Data is standardized, improved or derived e.g., profitability, scores
 - Promotes consistent analysis using common business rules
 - Reduces analysis “programming” since necessary information is produced prior e.g. calculating months on books, banding score values, computing household product counts
 - Increases understanding of the data
- C. Integrate - All of the required data is in one logical structure e.g., transaction, position, account, customer, household, product, instrument, branch
 - Simplifies data access because all data is located in one location
 - Reduces analysis time since all of the information can be retrieved from a single location through a single query

Review from Week 1: Data Analysis Methodology (cont.)

- D. Organize BI - Data is stored dimensionally
 - Simplifies analysis by providing an intuitive, business oriented data design
 - Enables pre-stored aggregations (cubes w/drill through to detail) to be easily developed and managed
- E. Organize A/DM – Data is stored in modeling specific data structure
 - This could be one de-normalized fact table with select dimensional information included in each row of data
 - Since data mining can only uncover patterns already present in the data, the sample should be large enough to contain significant information, yet small enough to process (dependent on resource capability)
- F. Explore - Search for anticipated relationships, unanticipated trends and anomalies:
 - Clustering discovers groups or structures in the data that are similar, beyond the structures known in the data
 - Classification generalizes a known structure to apply to new data, such as classifying a customer as a good or poor credit risk
- G. Document - Data definitions and transformation rules are documented and accessible
 - Able to understand the data and information gathered from the data

Review from Week 2:

Presentation - *best practices*

- ◆ Any inserts from Excel should be copy/pasted as a picture. There should be just one source of data. If that source of data is Excel, then keep it there. Imbedding objects can get tricky.
- ◆ If you insert charts or graphs from Excel, just insert the chart and/or graph. Exclude the menu bar and other surroundings.
- ◆ If you insert data from Excel, make sure to insert just the relevant cells and data, without background gridlines.
- ◆ With accompanying summary description – describe what the reader needs to know about the content that was presented, don't just copy/paste without a statement
- ◆ With formatting – uniform gridlines, wraps in cells, size the cells appropriately, size the chart/graph
- ◆ With readable sizing – no magnifying glasses but not 80 years old
- ◆ Without background gridlines – just turn it off for printing altogether
- ◆ Without dropdowns and filters that are Excel specific – if you include everything in Excel, might as well just share the Excel file

Review from Week 2:

Data in Excel - *best practices*

- ◆ Summary to the front – start with a “main or summary” sheet that explains content of the workbook e.g., table of content, versions, dates, etc. Also, consider including:
 - Summary of objectives or an introduction to the workbook
 - Key metadata/information needed to understand the analysis/data
 - Summary of findings
 - Updates to the workbook (if versioned)
- ◆ Data to the back – raw data sets, look ups, lists, etc. are references that should be in the back for referencing.
- ◆ Links to navigate – don’t make the reader spend half their time searching for references. Create links to datasets, if in another worksheet.
- ◆ Present the information first with a sentence or two before presenting the data – a reader should know what he/she is looking at before wondering what he/she is looking at
- ◆ Always sort the information. It’s about pattern. If you present chaos without pattern, it’s no information at all.
- ◆ Highlight derivation of data – always account for where the data came from and why you are using it for the analysis.

Week 3 Topic:

Submission Housekeeping

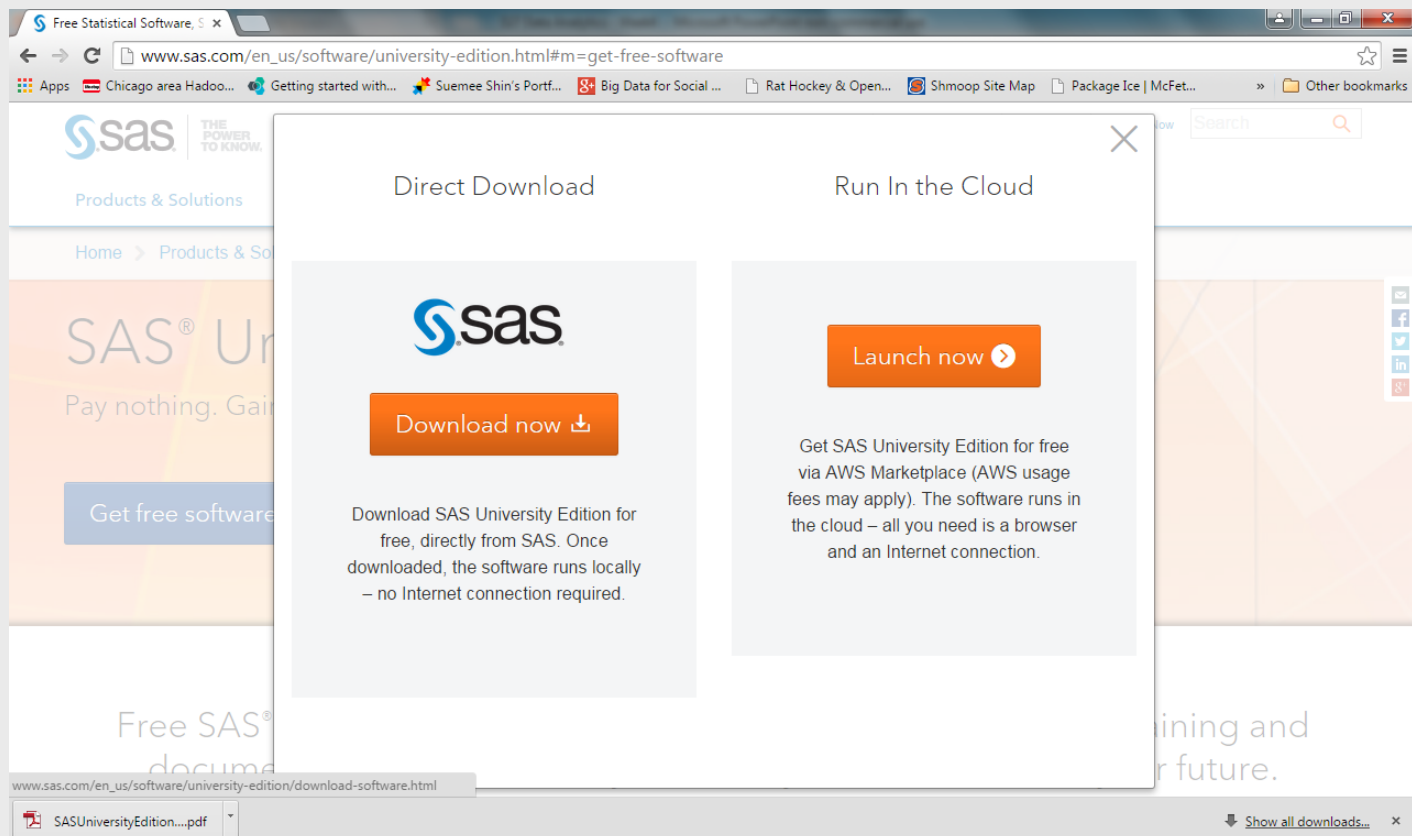
- ◆ Late submissions will not be graded
- ◆ If you submit multiple times, delete previous submissions. Will not be grading all submissions. Just the last.
- ◆ Do not zip files. Attach files individually
- ◆ Save with the first tab and first cell showing at all times in Excel. First cell shown for all tabs.
- ◆ Use following file name structure:
 <last name>_<first_name>_week<#>_<one word description of analysis>.ppt or xls

Week 3 Topic: Assignment 3

- 1) Read Chapter 2 of Data Smart – We will cover cluster analysis next week using the wine example in Chapter 2.
- 2) Case Study Part 1:
 - a) Collect datasets (no need to submit)
 - b) Create master workbook with all data. Merge 2011 ~ 2014 data together into one worksheet. Then, add transportation metrics, 311, and weather information into the same workbook.
 - c) Document assumptions & observations made while collecting and integrating data in a separate tab in (b) and map (hotlink) to attribute column as appropriate.
 - d) Document data validations, manipulations, & derivations in a separate tab in (b) and map (hotlink) to attribute column as appropriate.

Week 3 Topic: Installing SAS University Edition

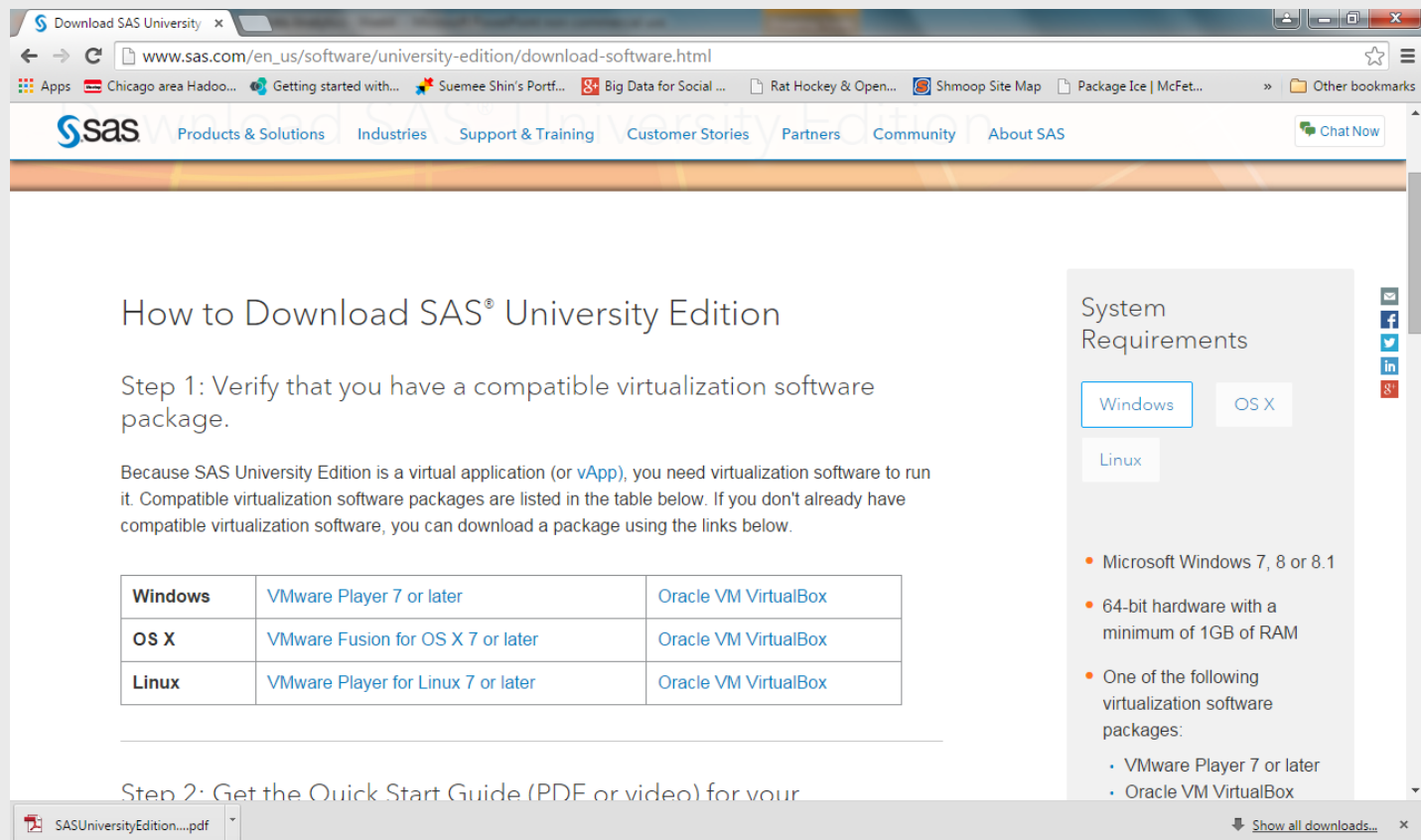
- ◆ Go to: http://www.sas.com/en_us/software/university-edition.html
- ◆ Create a student account. You will get free software and e-learning for one year.



Week 3 Topic:

Getting the right download file

- ◆ Select your system specific download files. Recommend using VirtualBox for Mac users.



The screenshot shows the SAS University Edition download page. The browser address bar displays www.sas.com/en_us/software/university-edition/download-software.html. The page title is "How to Download SAS® University Edition".

Step 1: Verify that you have a compatible virtualization software package.

Because SAS University Edition is a virtual application (or [vApp](#)), you need virtualization software to run it. Compatible virtualization software packages are listed in the table below. If you don't already have compatible virtualization software, you can download a package using the links below.

| | | |
|---------|--|--------------------------------------|
| Windows | VMware Player 7 or later | Oracle VM VirtualBox |
| OS X | VMware Fusion for OS X 7 or later | Oracle VM VirtualBox |
| Linux | VMware Player for Linux 7 or later | Oracle VM VirtualBox |

Step 2: Get the Quick Start Guide (PDF or video) for your

System Requirements

Windows OS X Linux

- Microsoft Windows 7, 8 or 8.1
- 64-bit hardware with a minimum of 1GB of RAM
- One of the following virtualization software packages:
 - VMware Player 7 or later
 - Oracle VM VirtualBox

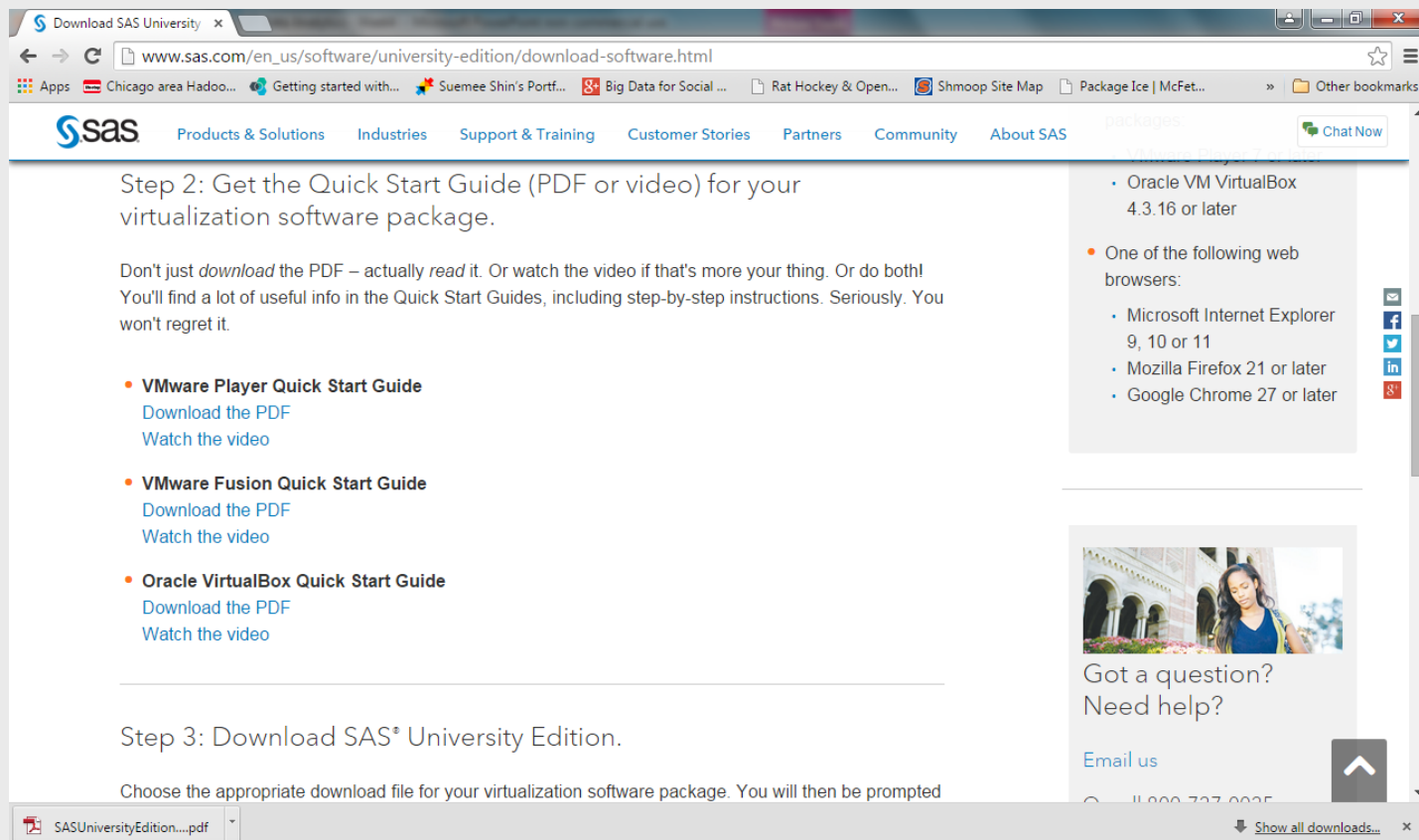
SASUniversityEdition....pdf

Show all downloads...

Week 3 Topic:

Follow installation directions

◆ Follow directions in the Quick Start Guide PDF



The screenshot shows a web browser window with the URL www.sas.com/en_us/software/university-edition/download-software.html. The page is titled "Download SAS University" and features the SAS logo and navigation links: Products & Solutions, Industries, Support & Training, Customer Stories, Partners, Community, and About SAS. A "Chat Now" button is also visible.

Step 2: Get the Quick Start Guide (PDF or video) for your virtualization software package.

Don't just *download* the PDF – actually *read* it. Or watch the video if that's more your thing. Or do both! You'll find a lot of useful info in the Quick Start Guides, including step-by-step instructions. Seriously. You won't regret it.

- **VMware Player Quick Start Guide**
[Download the PDF](#)
[Watch the video](#)
- **VMware Fusion Quick Start Guide**
[Download the PDF](#)
[Watch the video](#)
- **Oracle VirtualBox Quick Start Guide**
[Download the PDF](#)
[Watch the video](#)

Step 3: Download SAS® University Edition.

Choose the appropriate download file for your virtualization software package. You will then be prompted

On the right side, there is a "packages" section with the following content:

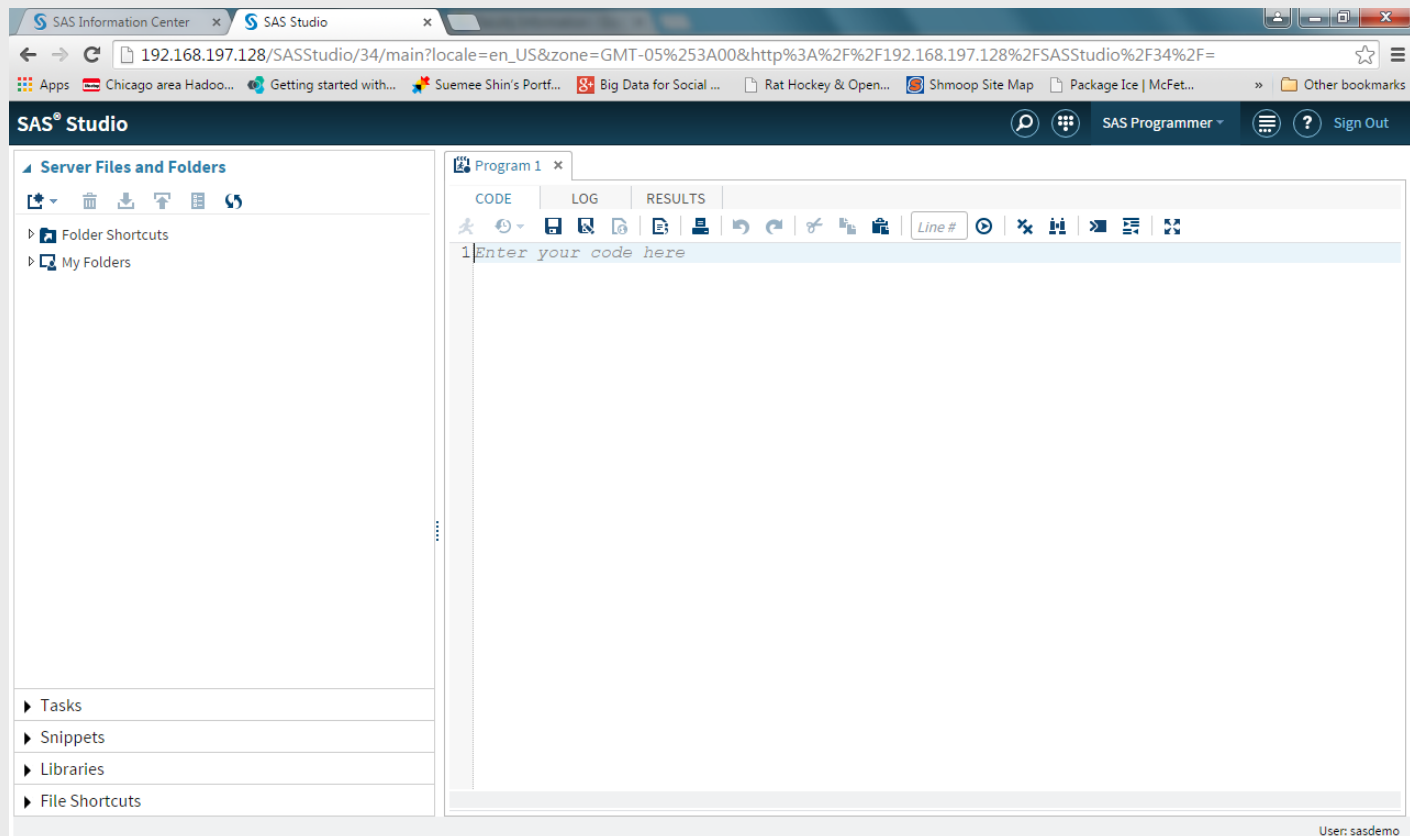
- Oracle VM VirtualBox 4.3.16 or later
- One of the following web browsers:
 - Microsoft Internet Explorer 9, 10 or 11
 - Mozilla Firefox 21 or later
 - Google Chrome 27 or later

Below this, there is a section titled "Got a question? Need help?" with a link to "Email us".

At the bottom of the browser window, a download bar shows a file named "SASUniversityEdition....pdf" with a dropdown arrow.

Week 3 Topic: SAS Studio – start window

- ◆ Get to the point of when you can open a start window:



Week 3 Topic:

SAS Programming I Essentials

- ◆ Start on SAS Programming I Essentials e-learning course as time allows. Ideally, we'll want to finish this course before Week 8:

