



# CODE-ALONG WITH US

Data Science Edition

**Data Viz**

Paul Tsagaroulis & Matt Albucher

# Data visualization

Introduction

## Graphs & charts

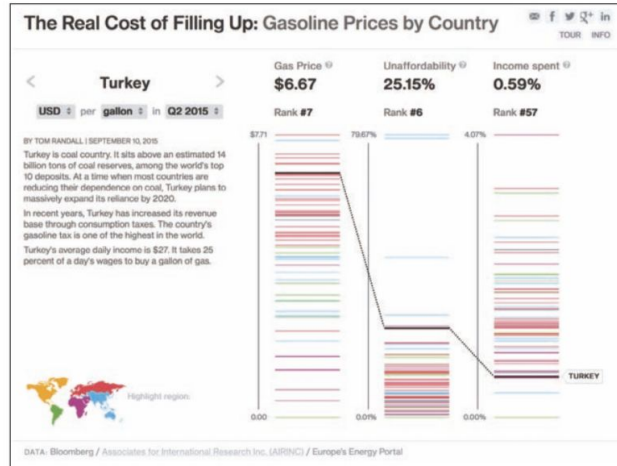
For comparison, distribution,  
composition, and relationships

## Visual analytics using

Open data demo



# Data visualization is everywhere



# Benefits of data visualization

---

**Having pictures in the analytic story is key to effective communication**

**We process graphs and images faster than text, which is why visualization is so effective**

# Visual information representation

---

## Graphical excellence

The greatest number of ideas, in the shortest time, using the least amount of ink, in the smallest space

## Visual integrity

Neither distort the underlying data nor create a false impression or interpretation of that data

## Data-ink ratio

Pay attention to how a visualization is compiled: Unnecessary elements should be removed

## Aesthetic elegance

Simplicity of the design evoking the complexity of the data clearly

# Gestalt principles of perception

## We organize what we see to make sense

*The whole is other than the sum of the parts*

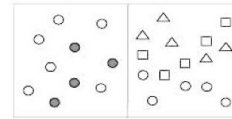
To understand visual perception in the underlying processes are organized to help us make sense of the world.

Proximity



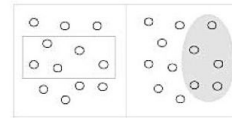
Objects are close together

Similarity



Objects sharing similar attributes (e.g., color or shape)

Enclosure



Objects with a boundary around them (e.g., formed by a line or area of common color)

Closure



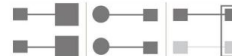
Open structures are perceived as closed, complete, and regular

Continuity



Objects aligned together or appear to be a continuation of one another

Connection



Objects that are connected (e.g., by a line)

# Preattentive visual properties

**Preattentive attributes are properties which are detected almost immediately without effort or extra processing by the brain**

Quick, effortless, and in parallel; without any attention being focused on the display

Color, form, movement, and space

Length, position, width, size, intensity/shade, orientation, shape, enclosure, position, mark

## Quantitative perception

Very precise

Length



Position

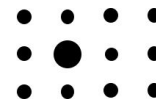


Not very precise

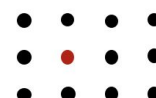
Width



Size

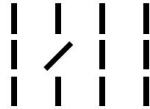


Intensity/shade

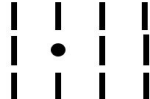


None

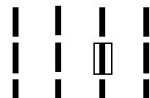
Orientation



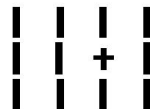
Shape



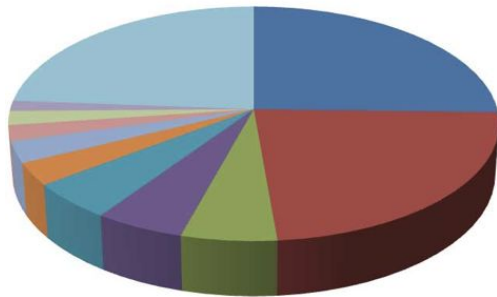
Enclosure



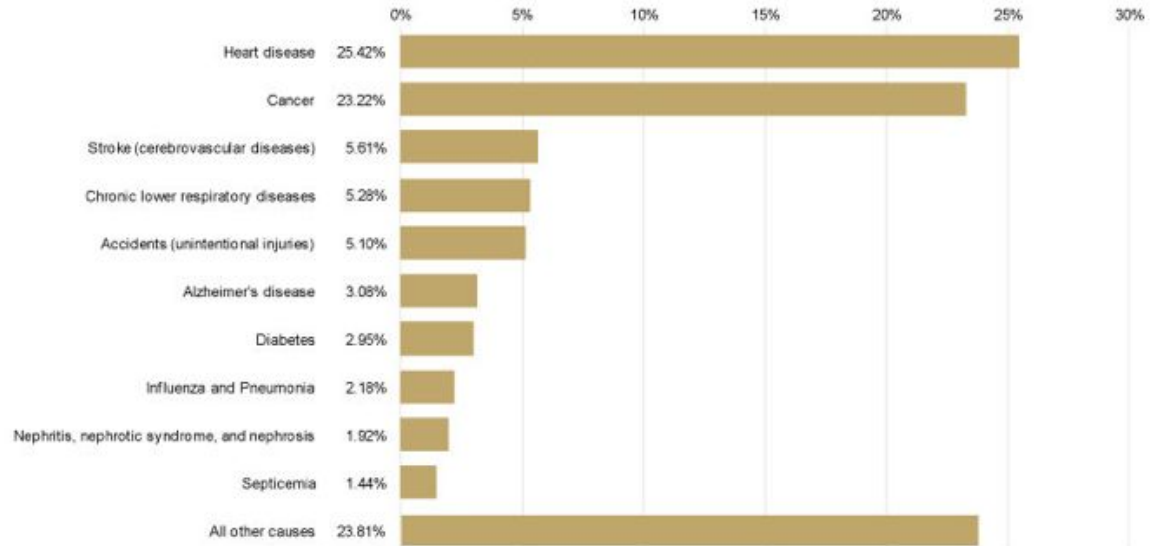
Mark



# Design principles in action



- Heart disease
- Cancer
- Stroke (cerebrovascular diseases)
- Chronic lower respiratory diseases
- Accidents (unintentional injuries)
- Alzheimer's disease
- Diabetes
- Influenza and Pneumonia
- Nephritis, nephrotic syndrome, and nephrosis
- Septicemia
- All other causes

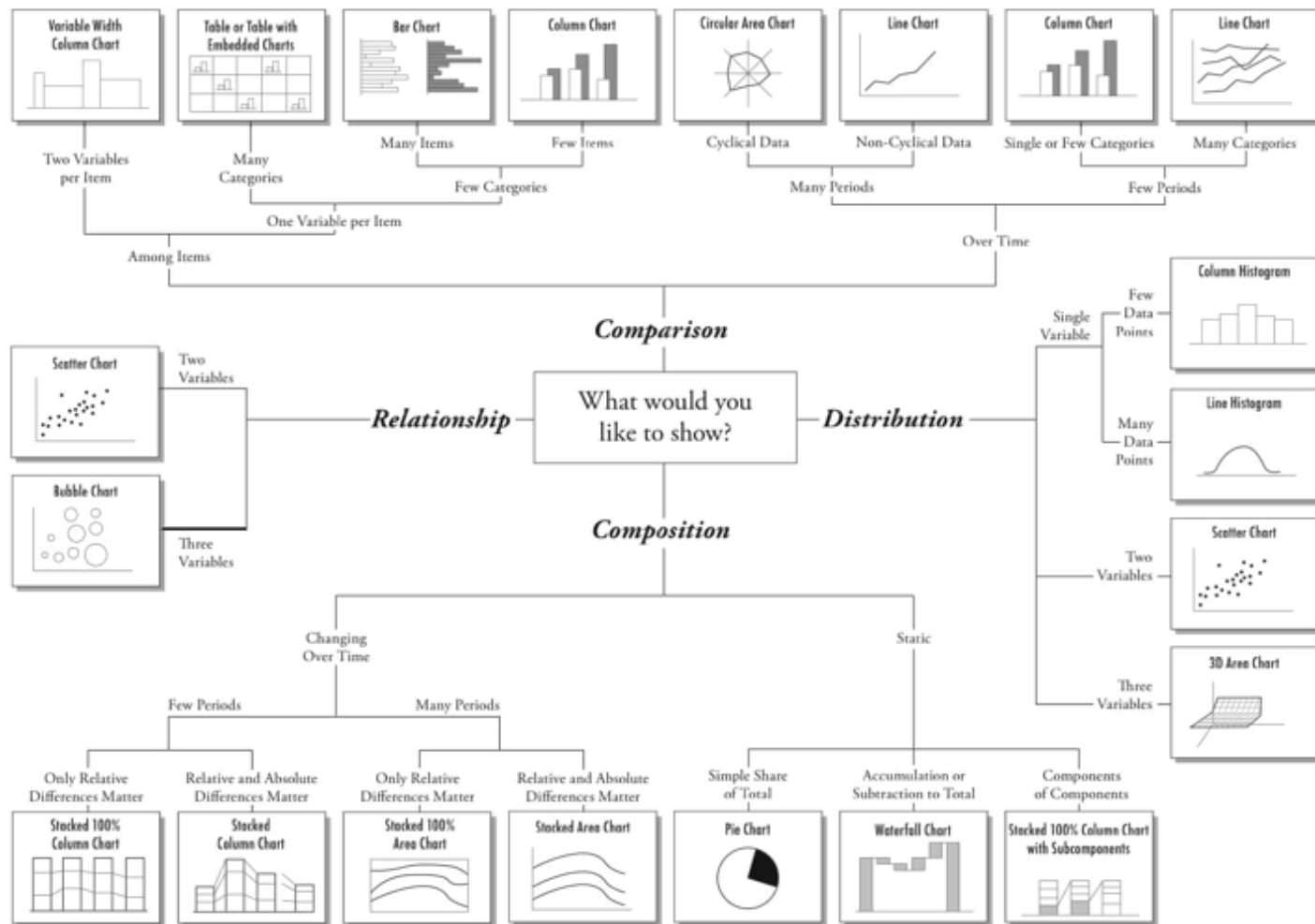




# Graphs & charts

# Chart Suggestions—A Thought-Starter

www.ExtremePresentation.com  
© 2009 A. Abela — a.v.abela@gmail.com



# Comparisons

---

**Bar/column graphs** can be used for analyzing data happening at a static point in time. Bar graphs are easily evaluated by examining the length of the bars; column charts are vertical.

**Line charts** can be used to display data or information that changes continuously, and allow us to compare how the data for different attributes changes over a period of time.

**Pie charts** are used to compare the parts of a whole. The graph is divided into several sectors, and the area in each sector shows the proportion they represent from the whole.

# Distributions

---

**Histograms** allow you to quickly assess shape, centering, and spread of distribution for a continuous data set. For categorical (nominal or ordinal) variables, bar charts are often used.

**Line charts** are the most frequently used charts and are used for continuous data sets. They are well accommodated for trend-based visualizations of data when the number of data points is very high.

# Composition

---

**Line charts** can accommodate the time component in an axis; it is useful for analyzing the trends of data over a period and useful for facilitating trend analyses.

**Dual-axis** is a special category of line charts. There are two independent axes that are layered on top of each other. These are useful when you have two measures that have different scales.

**Area chart** are the same as line charts however, the area below the plotted lines is filled with color.

# Relationships

---

**Scatter plots** show how much one variable is affected by another. The relationship between two variables is called their correlation.

**Bubble charts** are a variation of a scatter plot in which the data points are substituted with bubbles. The size of the bubbles can form a new dimension of the data.

**Line charts** can also be used for analyzing the relationships over a period of time.

# Visual analysis in R / Python with open data demo

# Using R / Python for Data Vis

---

## **A picture says more than a thousands words**

Visualized data can often be understood more efficiently and effectively than the raw numbers alone.

R and visualization are a perfect match.

Some must-see visualization packages are ggplot2, corrplot, ggvis, googleVis and rCharts.

## **Visualizations are an important criteria when choosing data analysis software.**

Although Python has some nice visualization libraries, such as Seaborn, Bokeh and Pygal, there are maybe too many options to choose from.

CRAN hosts an exponentially growing number of data visualization packages posted by worldwide users for a wide array of visualization applications.



# Data | GSA Open Technology

---

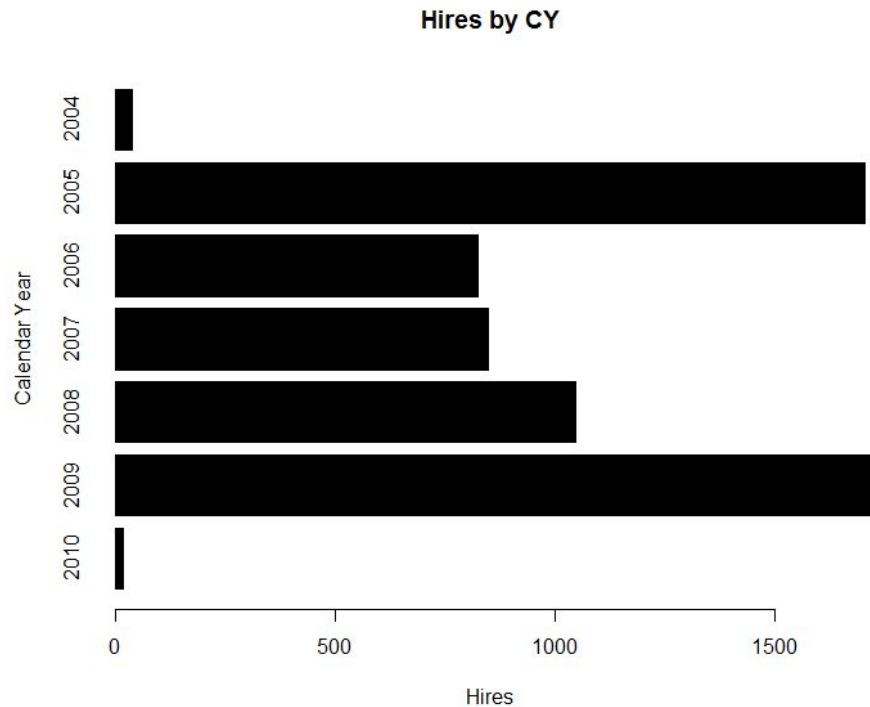
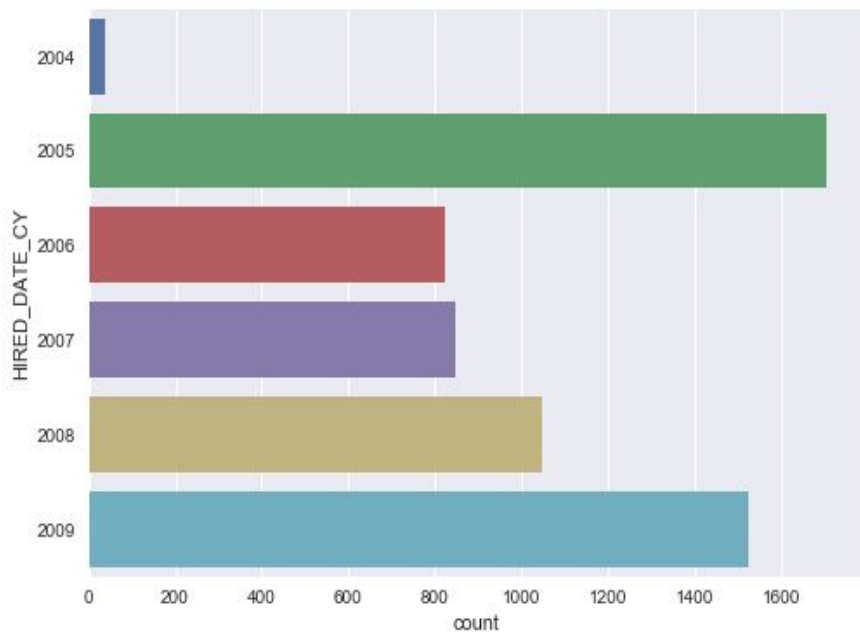
## Time to Hire

This dataset represents time taken to hire a GSA employee from the internal request to hire through the entry on duty of the of the selected individual.

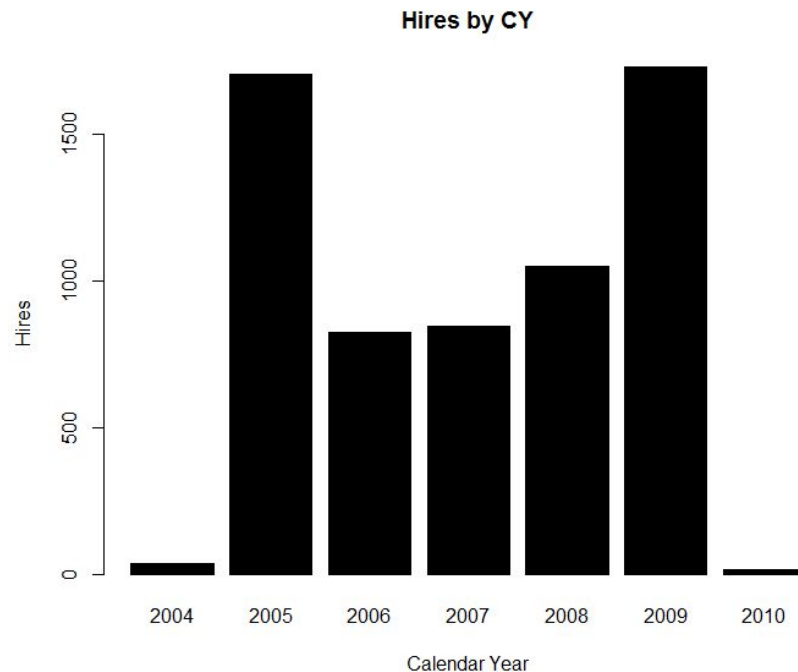
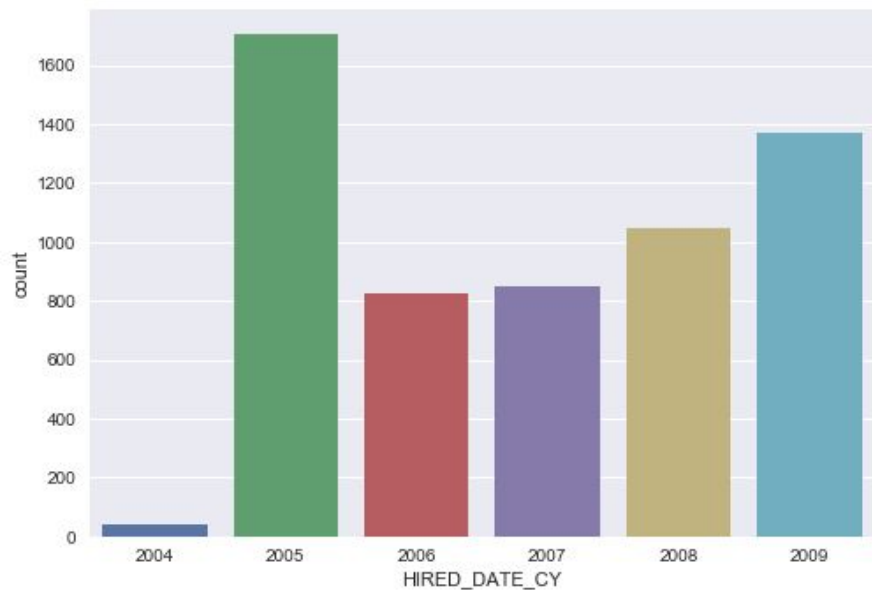
[View Project](#)

<https://open.gsa.gov/data/>

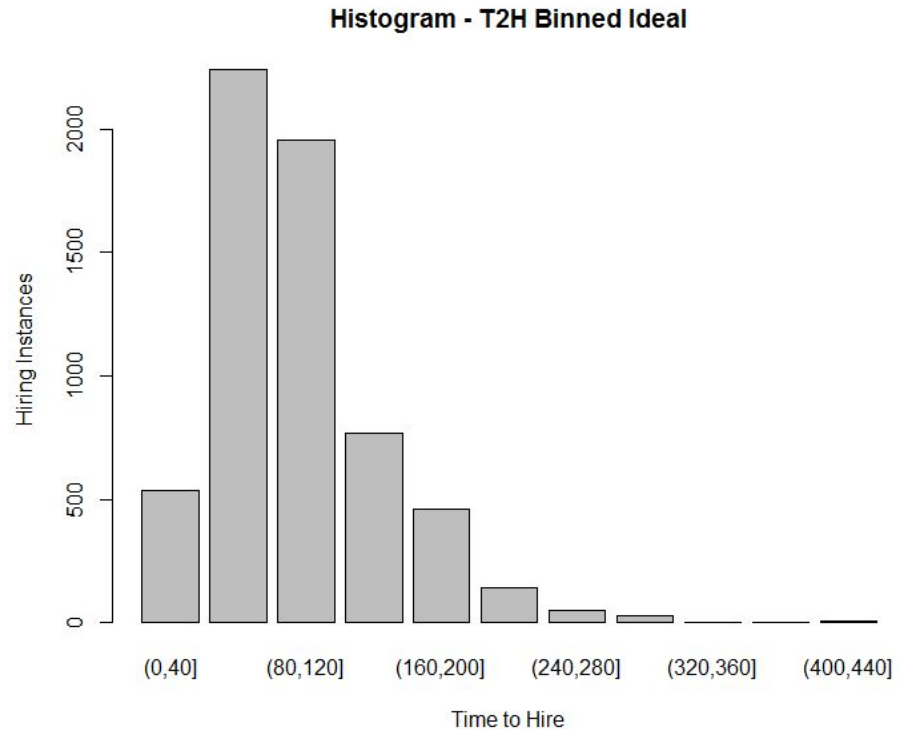
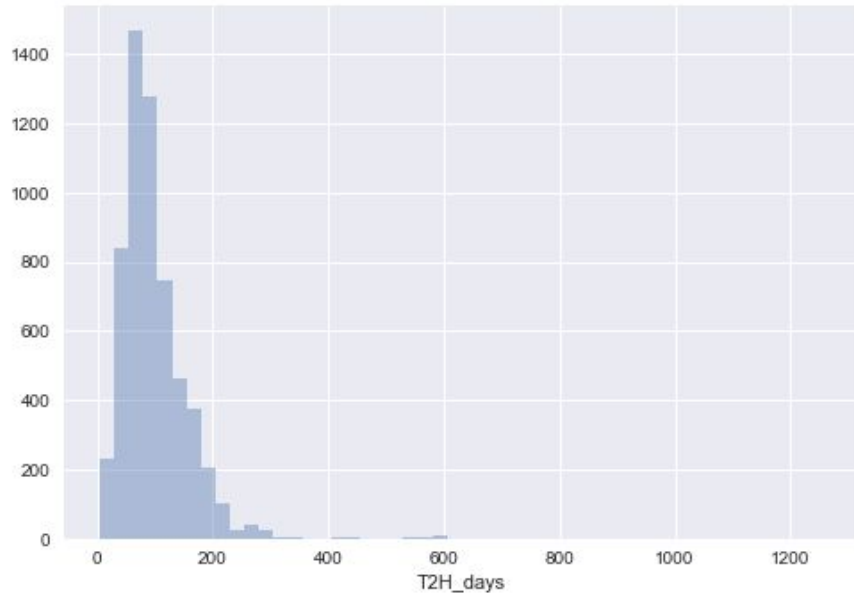
# Comparisons: Column graph



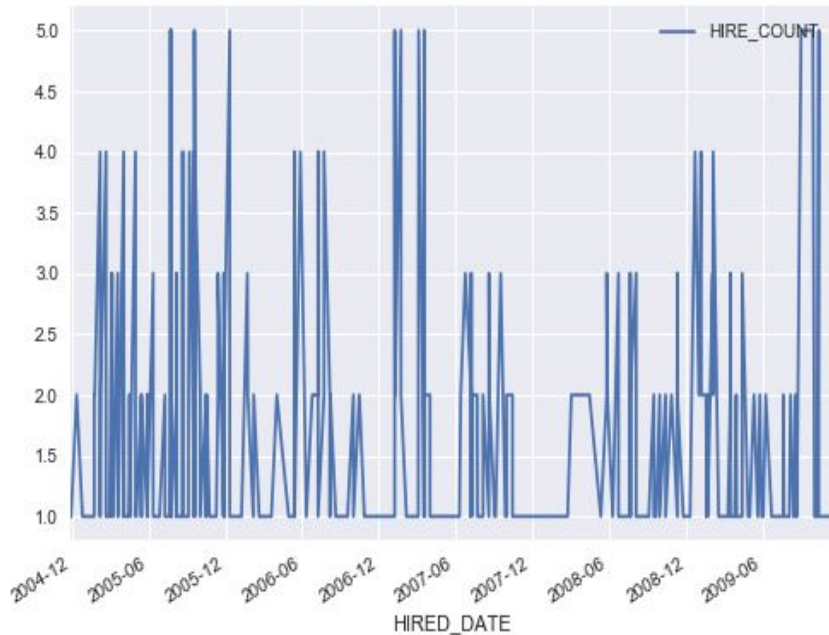
# Comparisons: Bar graph



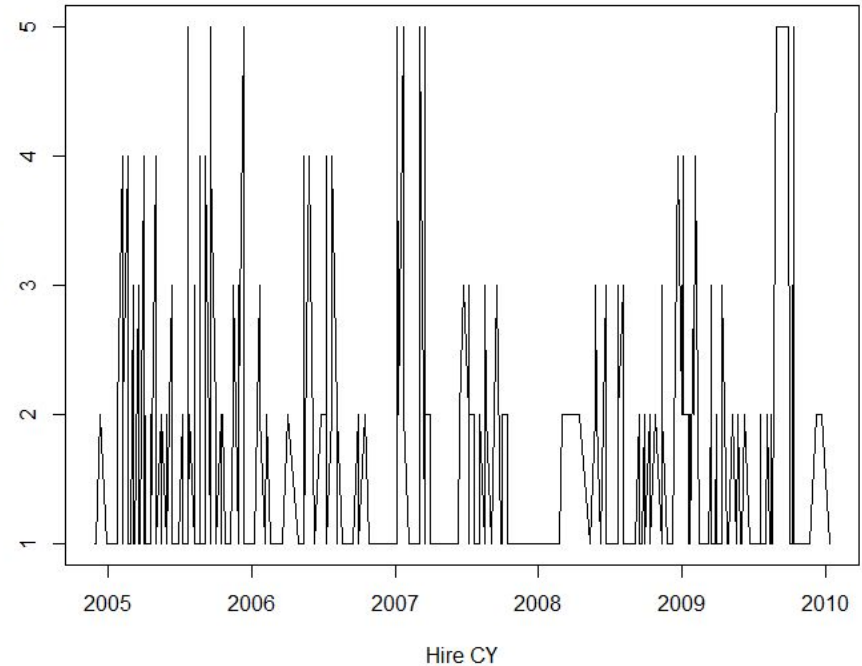
# Histogram: Time to hire measure



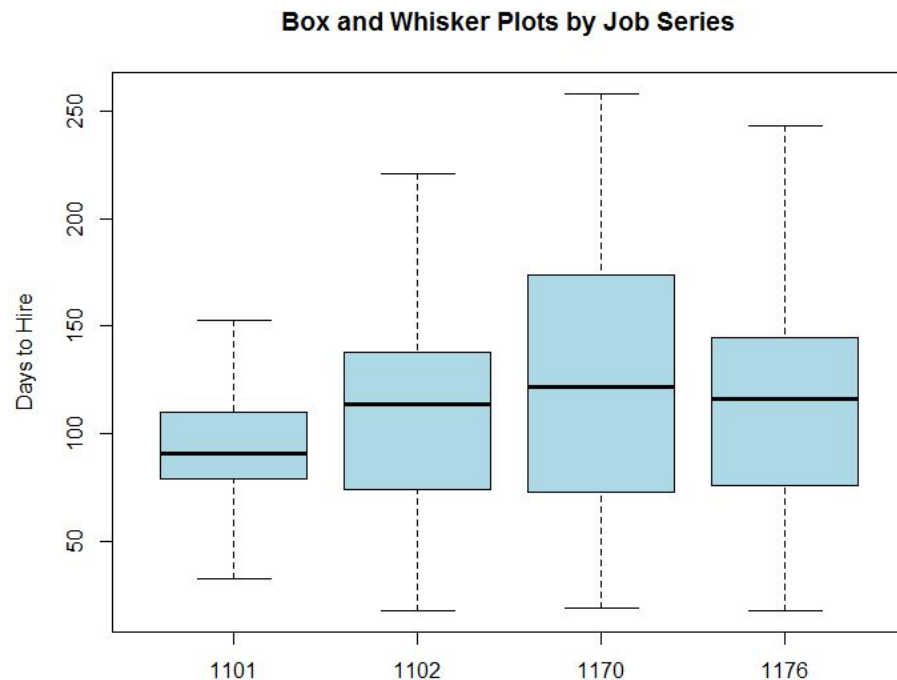
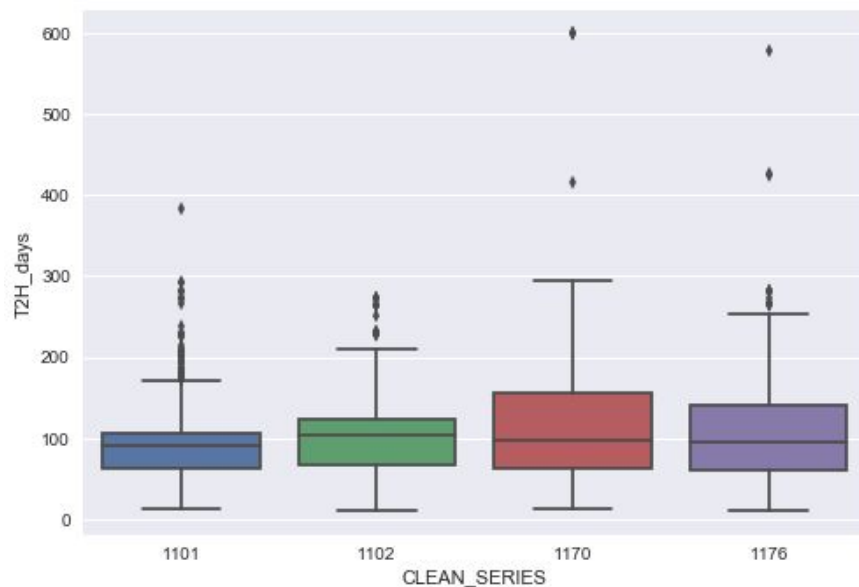
# Comparison: Line chart



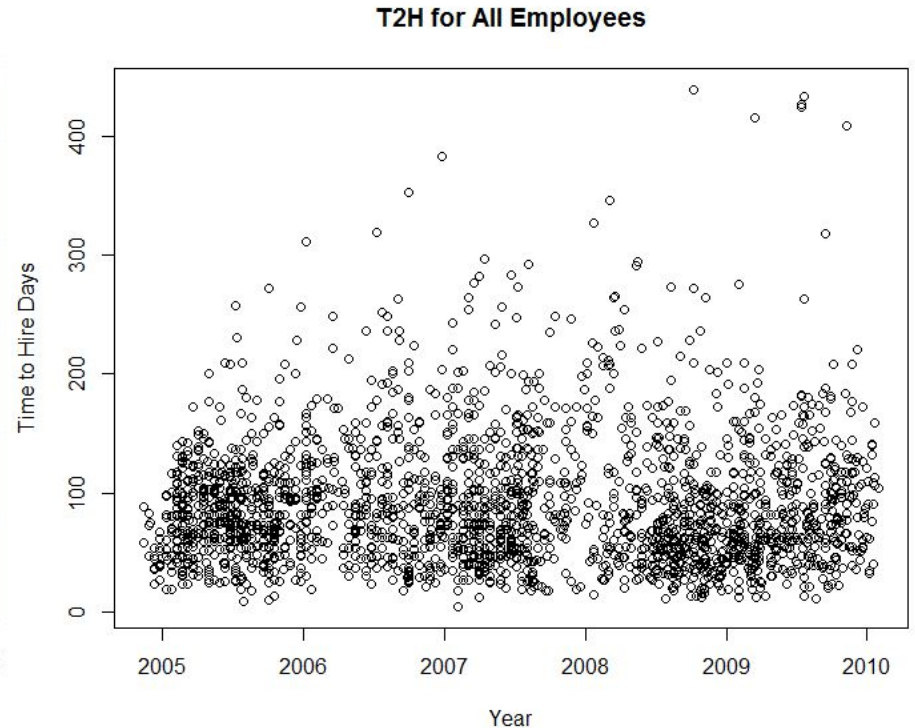
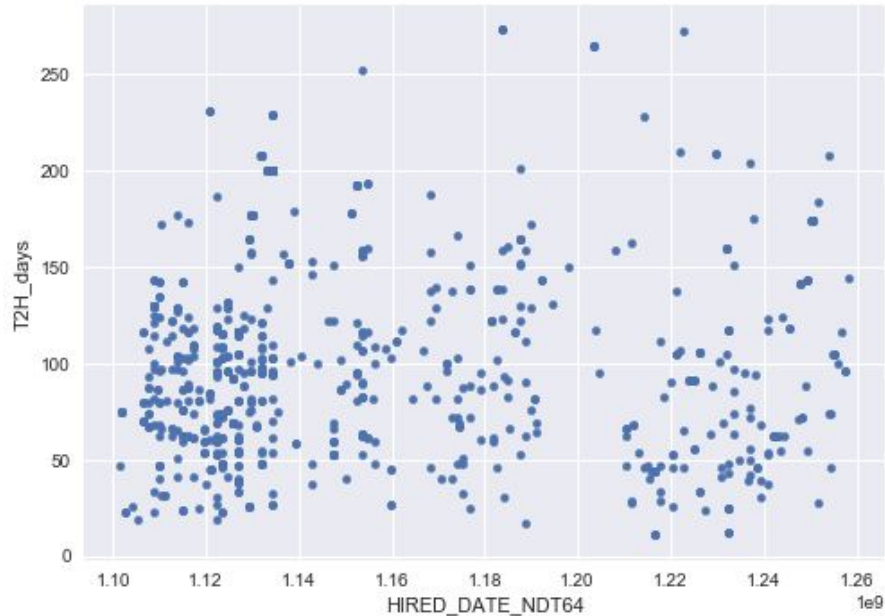
Hire Count over Time for 1102s - Line Graph



# Box & whisker: T2H by job series



# Scatterplot: T2H vs Hire Date



# Resources

## Python

[matplotlib.org/gallery](https://matplotlib.org/gallery)

[seaborn.pydata.org](https://seaborn.pydata.org)

[pbpython.com/simple-graphing-pandas.html](https://pbpython.com/simple-graphing-pandas.html)

[github.com/ukgovdatascience/Python-for-Analysts](https://github.com/ukgovdatascience/Python-for-Analysts)

## R

[ggplot2.org](https://ggplot2.org)

[ggplot2.tidyverse.org/reference](https://ggplot2.tidyverse.org/reference)

## Data visualization links

[google.com/site/showmethedatasessionnotes](https://google.com/site/showmethedatasessionnotes)

[github.com/GSA/training-pathway-data-practitioner/tree/master/specialization-data-visualization](https://github.com/GSA/training-pathway-data-practitioner/tree/master/specialization-data-visualization)



