February 22nd 2021

Exam, Presentation, Guest Speaker, Naive Bayes, Linear Regression, Clustering

Exam

What to expect (again)

Let me know when you want to take it (72 hours)

March Project ~ 7-10 Minute Presentation

WHAT IS EXPECTED

- Choose GREAT data
  - GREAT == You care about it
  - GREAT == Has {appropriate, interesting, workable} FEATURES
  - Data is robust in FEATURES or Volume
- Determine INTERESTING *Business Questions*
  - Set out to look for something you care about or find intriguing
  - ITERATE! Use initial findings to go down the rabbit hole
- Formulate INCREDIBLE conclusions and insights
  - Visual & Qualitative TAKEAWAYS - you actually think others should **‘care and share’**
- Spend time writing GOOD CODE (Ask Questions)
  - Don’t take shortcuts on quality non-code work, because you didn’t put the time in
  - Create incredible visuals with key insights, means setting up a good coding ‘test bed’

March Project - You are presenting to The Board

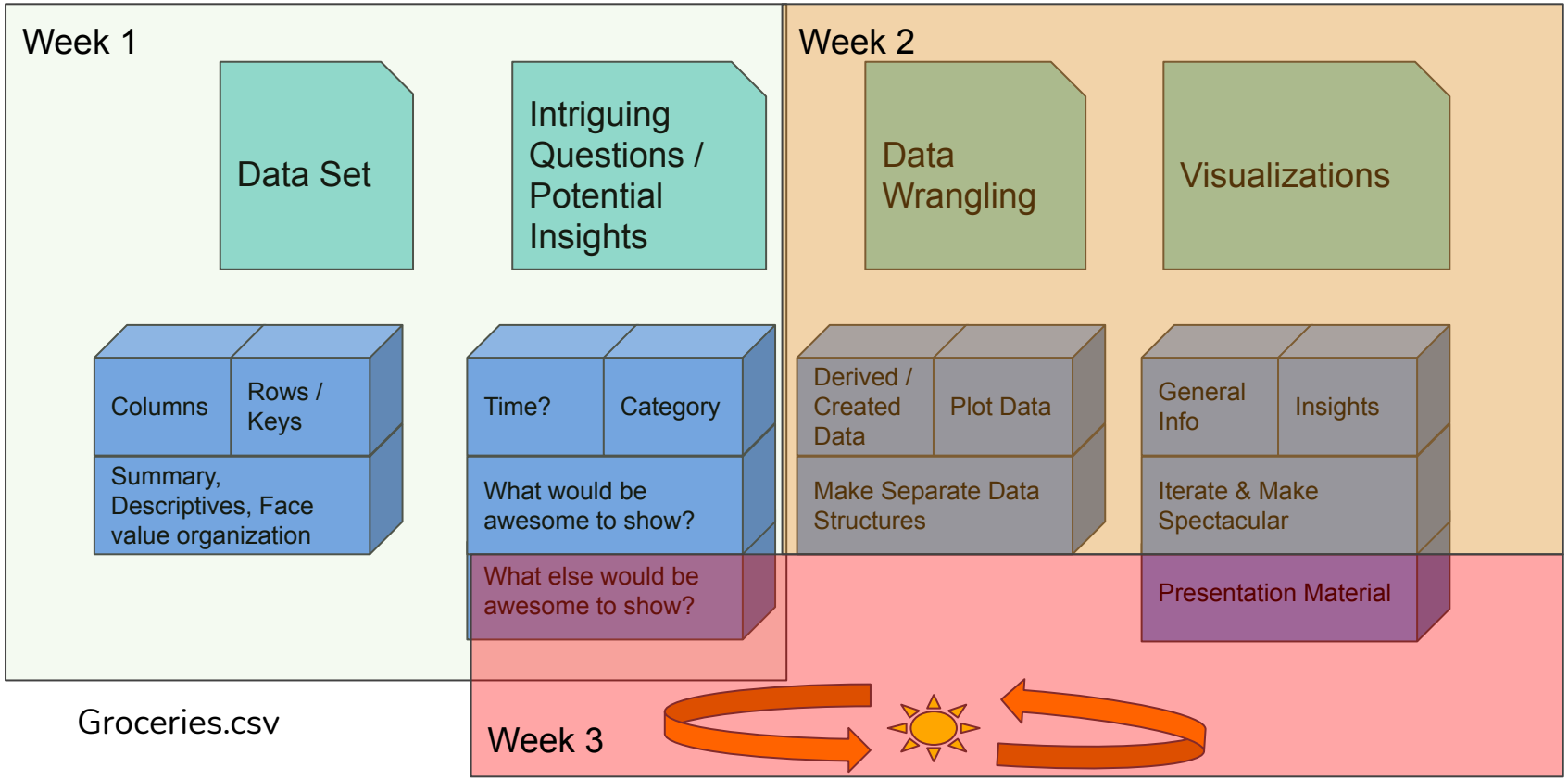
A SUGGESTED AGENDA FOR YOU TO FOLLOW

- Choose GREAT data, Determine INTERESTING *Business Questions*, & Sharpening HW
  - Week of Feb 15-21
- Data Munging, Wrangling, & Exam
  - Week of Feb 22 - Feb 28
  - ITERATE! Weeks of March 1 - 15
- Formulate INCREDIBLE conclusions and insights - Steve's Presentation
  - Weeks of March 8 -15
- Write CODE - Let me reiterate everything...
  - Do your Exploratory Data Analysis in FEB
  - Set yourself up to do A LOT of interesting discovery, try different segmentations, insights, correlations, visual inspection, ITERATE March 1-7+
  - Produce visualizations, stories, presentation talk track and presentation March 8-15

Subjectivity

PTS? TBD.	Low	Average	High	X Factor
Coding	Straightforward dataset, little to no derivative data, minimal ‘munging’ attempts	Some *derivative data created*, wrangling attempted and succeeded, data structures used	Several creative data slicing and filtering methods, m*uch of the used data was derived or created*	Did you have to? Comparison? Roadblocks conquered?
Math / Algorithms	Count, sum, max, min, etc.	Trends, rate of change, %, mean	Probability, covariance, standard deviation, Naive Bayes, Regression, Clustering	Did you use these ‘correctly’ and ‘interestingly’ or just use them to use them?
Analysis	Tell us about the above	Opinions and insights on why, call outs to external events or knowledge	Correlated nuggets of information coming from *derived data creation*, external callouts, attempting to answer why, how, potential cause and effects	Is this information you would not have guessed was the case anyway? Why is this ah ha? Shareable? Comparable?
X Factor	Basic plots of the above, presentation skills	The right plots, appropriate and informative information within, presentation skills	Incredibly insightful and visually appealing plots, incredible presentation techniques	**Comparison**. Take away story? Was there a surprise?

Project Planning - My Week 1+



March Project - Last Notes

Algorithms / Concepts Covered by March 8th

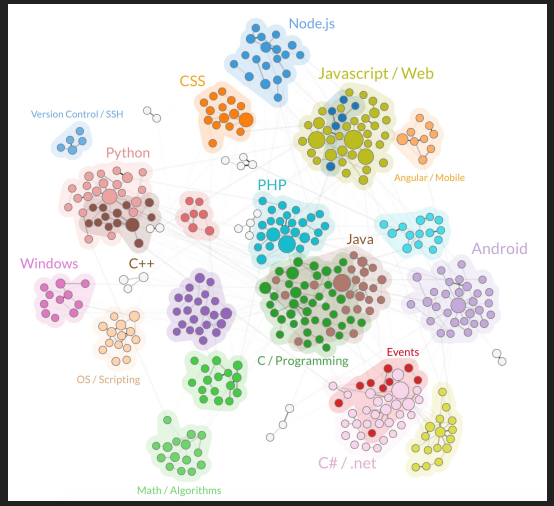
Clustering and Regression

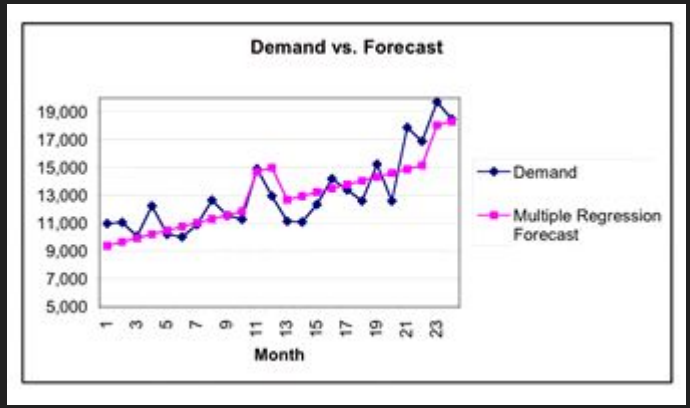
Presentations by np.random.choice(students)

- March 15 (submissions)...some March 22

NOTES:

1. Really get into what the data/columns mean
2. Enumerate how you could roll up the data
3. Location, time, loop through each column…
4. Off the top of my head...12 Hours

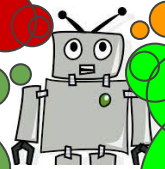




Project Tips - My Week 1+

1-2 Min - introduce the dataset, columns, describe, what did you set out to do

1-2 Min - summary stats and visuals (2-4), show the audience what the data set gives us out of the box...counts, sums, high level stuff and explanations of top/bottom 10's



1-2 Min - maybe jump over to Jupyter Notebook (or screenshots) and show them some of the messy wrangling I did, but more importantly, describe why I had to do it...what is in store for them after this...queue dramatic music

1-2 Min - Wow them with 5+ graphs of insights, since my data naturally screams '..., it's perfect for an this type of algorithm.

Conclusion & Questions

Jupyter Notebook - Data Munging Tips!

If you draw up the right game plan - “I know what I want to do…but the coding is grr…”
You are on the right track!

60%…maybe 70%…. Will be spent on data wrangling, data munging, data wrestling…

Ask for help - Poisson Distribution as we move into March

If you do these things…the visuals…the iterations on all of this…will come very easily

Build your presentation deck outline early / often … make it effective not in replacement for…

Presentation DO NOT’s - which may conflict with other teachings and I am OK with that…
Bullets are boring? Words on a slide vs. spoken? Animation?

5 Minute Sidetracks

TEDx Talk

AI Article...thoughts, questions, comments

https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

Tide case study from Ogilvy - 5 More Minutes

How might you be asked to use data?

Finding unbiased nuggets or hunting for a story

What is inherent to ‘slicing data up’

Categorical columns!! Look here first.

Numerical columns which you feel deserve buckets - Working Class?

Groupings == Segmentation

	TEAM	CONF	G	W	ADJOE	ADJDE	BARTHAG	EFG_O	EFG_D	TOR	...	FTRD	2P_O	2P_D	3P_O	3P_D	ADJ_T	WAB	POSTSEASON	SEED
0	North Carolina	ACC	40	33	123.3	94.9	0.9531	52.6	48.1	15.4	...	30.4	53.9	44.6	32.7	36.2	71.7	8.6	2ND	1.0
1	Wisconsin	B10	40	36	129.1	93.6	0.9758	54.8	47.7	12.4	...	22.4	54.8	44.7	36.5	37.5	59.3	11.3	2ND	1.0
2	Michigan	B10	40	33	114.4	90.4	0.9375	53.9	47.7	14.0	...	30.0	54.7	46.8	35.2	33.2	65.9	6.9	2ND	3.0
3	Texas Tech	B12	38	31	115.2	85.2	0.9696	53.5	43.0	17.7	...	36.6	52.8	41.9	36.5	29.7	67.5	7.0	2ND	3.0
4	Gonzaga	WCC	39	37	117.8	86.3	0.9728	56.6	41.1	16.2	...	26.9	56.3	40.0	38.2	29.0	71.5	7.7	2ND	1.0

What the real world looks like

For a data scientist - mostly the same as last slide

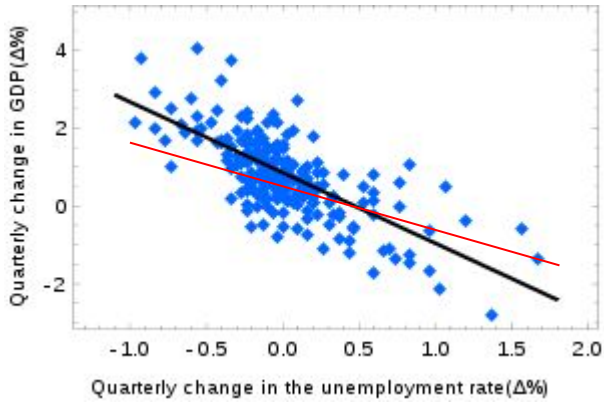
For others - directed request, usually a boondoggle that you need to shape

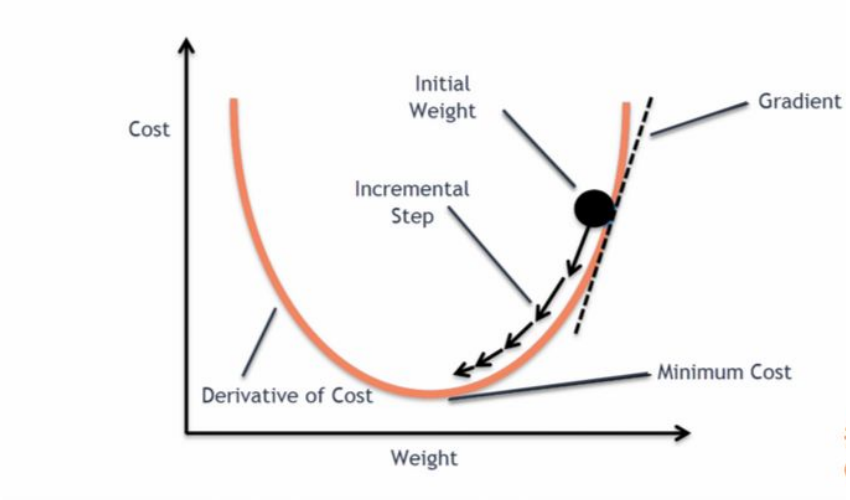
Groupings == Segmentation

	TEAM	CONF	G	W	ADJOE	ADJDE	BARTHAG	EFG_O	EFG_D	TOR	...	FTRD	2P_O	2P_D	3P_O	3P_D	ADJ_T	WAB	POSTSEASON	SEED
0	North Carolina	ACC	40	33	123.3	94.9	0.9531	52.6	48.1	15.4	...	30.4	53.9	44.6	32.7	36.2	71.7	8.6	2ND	1.0
1	Wisconsin	B10	40	36	129.1	93.6	0.9758	54.8	47.7	12.4	...	22.4	54.8	44.7	36.5	37.5	59.3	11.3	2ND	1.0
2	Michigan	B10	40	33	114.4	90.4	0.9375	53.9	47.7	14.0	...	30.0	54.7	46.8	35.2	33.2	65.9	6.9	2ND	3.0
3	Texas Tech	B12	38	31	115.2	85.2	0.9696	53.5	43.0	17.7	...	36.6	52.8	41.9	36.5	29.7	67.5	7.0	2ND	3.0
4	Gonzaga	WCC	39	37	117.8	86.3	0.9728	56.6	41.1	16.2	...	26.9	56.3	40.0	38.2	29.0	71.5	7.7	2ND	1.0

Topics - Linear Regression

$$\text{Find } \min_{\alpha,\,\beta} Q(\alpha,\beta), \quad \text{for } Q(\alpha,\beta) = \sum^{n} \hat{\varepsilon}_i^{\,2} = \sum^{n} (y_i - \alpha - \beta x_i)^2$$





Remember - the algorithm for our algorithms is also an algorithm...or something like that...

```
In [24]: from sklearn.svm import SVC

         model_name = 'Kernel SVM Classifier'

         svmClassifier = SVC(kernel='rbf', gamma= 'auto')

         svm_model = Pipeline(steps=[('preprocessor', preprocessorForFeatures),('classifier', svmClassifier)])

         svm_model.fit(X_train,y_train)

         y_pred_svm = svm_model.predict(X_test)
```

```
In [18]: from sklearn.linear_model import LogisticRegression

         model_name = "Logistic Regression Classifier"

         logisticRegressionClassifier = LogisticRegression(random_state=0,multi_class='auto',solver='lbfgs',max_iter=1000)

         lrc_model = Pipeline(steps=[('preprocessor', preprocessorForCategoricalColumns),
                                     ('classifier', logisticRegressionClassifier)])

         lrc_model.fit(X_train,y_train)

         y_pred_lrc = lrc_model.predict(X_test)
```

```
In [217]: from sklearn.naive_bayes import GaussianNB

In [218]: classifier = GaussianNB()

In [226]: features = zip(data.W[:1600], data.ADJOE[:1600], data.WAB[:1600])
          test = zip(data.W[1600:], data.ADJOE[1600:], data.WAB[1600:])
          #classifier.fit(features, data.SEED[:1600])
          classifier.fit(np.array(data.W[:1600]).reshape(-1,1), data.SEED[:1600].astype(int))

Out[226]: GaussianNB(priors=None, var_smoothing=1e-09)

In [222]: preds = classifier.predict(test)
          print(preds)
```

Topics

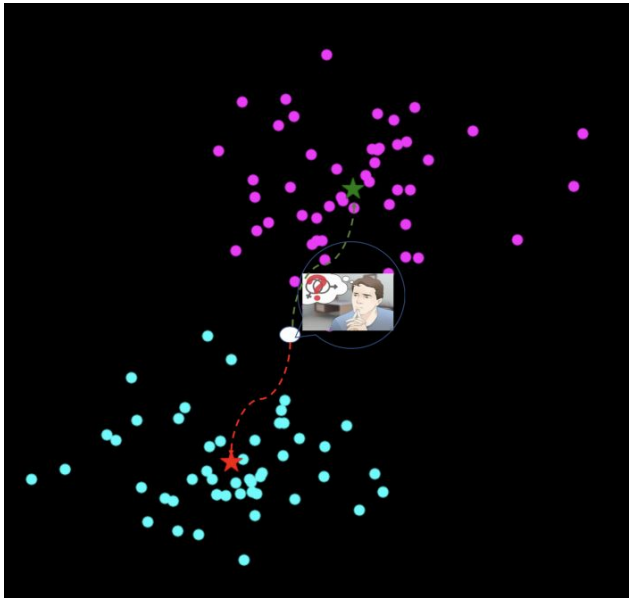
K-Means Clustering

Choose K Clusters...

Find the ERROR

Move Everyone to their closest cluster...

REPEAT until nothing changes!



Jupyter Notebook

Regression

Walkthrough

Coding