Exploration - Now I Know Game Results

Now I Know

We exploration to discover features that drive the target
You have to explore bivariate and multivariate after splitting
Only only explore on train
Heatmaps only for continuous variables
start by forming your initial questions + hypotheses

1. Ask questions in English about your data (what's driving y)
2. Analyze variables to answer that question
3. Write down your takeaways in plain English (no buzzwords)
   hypothesis testing when visualization isn't clear
   encode the data before splitting
   Recommend exploring un-encoded categorical values
   Recommend exploring un-scaled data
   - because exploration viz + tests is for people
   - People read "red, green, blue" a lot easier than the encoded version
   - People understand "degrees and feet" not scaled versions of those
   - Approaches:
     - for each column like payment_type, make an encoded_payement_type column on that same datafrmae
     - another approach: keep your original dataframes, but make encoded + scaled versions of each dataframe

I think I've got it, but I'm not too sure about

Confused about a confusion matrix
- Only use confusion matrix on classification models (b/c classification models come from having a cateogry target)
- Real talk: know your confusion matrix terminology for interview questions. On the job, you use the classifcation_report function from sklearn.

Which hypothesis test to use in which situation, Am I using the right features w/ the right right statistical test

- Categorical to categorical to check for independence is chi-squared
- Numeric to numeric? Use correlation (pearson for linear w/ normality or spearman for non-normal distributions)
- Numeric to categorical? Use a t-test (w/ normal) or mann-whitney (non)
    - If we're comparing a numeric value like monthly_charges between two populations (churn and stay), this is a independent t-test
    - If we're comparing one subgroup like senior_citizen DSL customers to the entire population, that's a 1 sample t-test
    - T-test and mann-whitney are only comparing 1 "binary" with one numeric.
    - What if we need to compare monthly_charges DSL vs Fiber vs. no-internet or 4 values or more? (ANOVA test) and if your numeric column doesn't satisify the assumptions of ANOVA (normality), then kruskal-wallace to the rescue
    - 1sample t-test and independent t-tests can both be either one tailed or two-tailed
        - 2 tailed test is checking for a difference in any direction between the numeric values (monthly_charges are different for churned vs. non-churned)
        - 1 tailed test checks for directionality
- Whether if we discover evidence for a relationship or not, not finding something is a finding. Like gender not mattering with churn.
- What about a question like: DSL customers more likely to churn or Fiber? (Since this is a chi-squared but chi-squared doesn't give us more or less comparisons)
    - We can answer this question with pandas
    - Make a subset df for for DSL customers then check the proportion of churned customers. Then make a subset df of Fiber customers then check the proportions of churned customers.

what visuals to use in each situation, Which visualization answers the question best?

- Always histogram at your target variable and any variables that you end up needing to use in a hypothesis test.
- Initial exploration of numeric to numeric, use a pairplot to go fishing looking for linear relationships, polynomial (curvy) relationships, or no linear or curvy relationship.
- If you have a binary target variable, you can use a bar plot with the target on the y axis to show proportion
- Numeric to numeric? Scattplot. Line plots are neat if you need to connect each successive datapoint (think stock chart). Scatterplots show you actual data points
- https://seaborn.pydata.org/examples/index.html provides visual examples
- Which categorical plots to use? https://seaborn.pydata.org/tutorial/categorical.html
- More continuous options https://seaborn.pydata.org/tutorial/relational.html

which tests assume a normal distribution (see above)
recording takeaways, recording takewaways - write it as soon as you discover it, not "I'll document first thing in the morning"
confused about null hypothesis *journal as you go*

Am I leaking my data? (not if you stick to train, so stick to train, only fit your model on train, tune your hyperparameters on train)

Do I need to finish all exploration before I start doing preparation? We may discover that combining featgures or removing features will be helpful in explore and need to back to the prep script to do something with them. If you discover something you missed or need to change,

Do I need to finish every single piece of prepartion before exploring? No. We can explore our target and handful of prepped variables to get some forward momentum.

Have I explored enough? (when is enough enough?)
- Explore can be a maze you get lost in.
- If time is precious, #1 goal is moving forward w/ an understqanding of driver of your target.
- Scenario setup: If I get a surpise extra week on a project, I'll put that into explore > modeling.
- If you're on the spot and need to present earlier than expected (surprise!), 2-3 takeaways is all you need.
- "The process of making a second draft is the process of making it look like you knew what you were doing all along" - Neil Gaiman
- Focus on cost vs. benefit of your time.

While modeling, can I come back to exploration? Yes. Get your MVP model out the door. You MVP model comes from your MVP exploration. Once you have that MVP of the entire presentation (explore + model + presentation slide)

Is it clear enough for the audience? (Avoid buzzwords, avoid sounding too mathy, be direct, simplify)

When is a normal-ish distribution OK vs. too non-normal?

- If your historam looks like a plot of of bitcoin price, it ain't normal
- If your histogram looks like an italicized bell curve
  Have I underfit or overfit my model? (can I make my improve modeling through exploration), then it's normal enough for a t-test.

automation w/ GridSearch? see https://ds.codeup.com/advanced-topics/cross-validation/#grid-search-cv.

Am I gettting the correct data? (Always check to make sure the data you think you're dealing with is actually the data you're dealing with). Example: if you change your SQL query but your wrangle script was checking for a local .csv and you don't the the results of the SQL query change?

Do i need to do stats tests even if my visual is clear? Nope.

We say we want to ask questions + answer them, but how do we ensure effective question formation? Are my questions too basic and how do I make sure that the questions get me where I need to go?

- Initial question formation can come from stakeholders where we need to refine their problem statement. Example: "who are our customers?" from the stakeholder. We might answer with who the best customers are or do some sort of grouping/segmentation
    - Other initial questions can come from hunches or things "everybody already knows" but check it!
        - "Obvious" things or even super simple, basic questions are worth looking into to be sure.

Sometimes one question can lead to 100 other questions. Now what? How to stop yourself from over-exploring?
  - Comes downs to cost-benefit of time + deadlines.
  - If you have a meeting Friday, 2-3 takeaways will beat a perfect analysis of everything.

How to summarize your findings to make a good story for your audience?

- We can use established narrative patterns like "Man in a hole" or a 3 act play: In the beginning, we found that there's LOTS of churn. When we looked into the data we discovered X was churning a bunch and Y wasn't. Now, let's talk about how to find more Y customers. Beginning -> Middle -> End
- Continuously remind yoruself through the project and written down in your notes/markdown/slides, "why does this matter" and "Why should you care"
- Less is more. A handful of powerful takeaways beats 1000 takeaways that don't really matter.