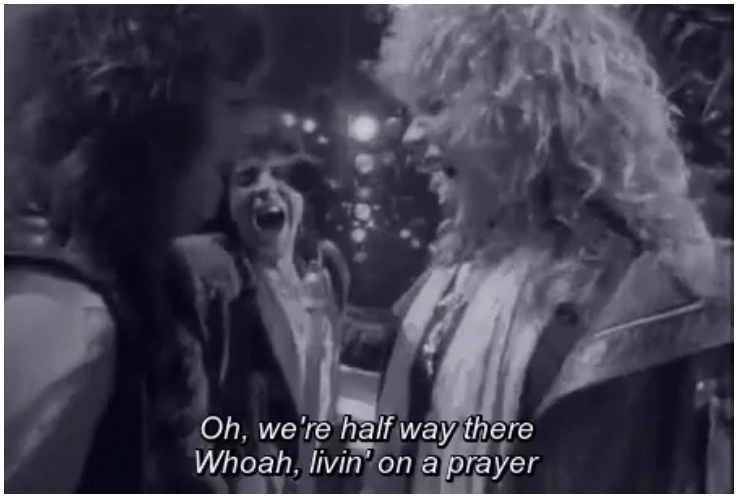
Phase 2 Project Description

Another module down - you're almost half way there!



All that remains in Phase 2 is to put your newfound data science skills to use with a large project!

In this project description, we will cover:

- Project Overview: the project goal, audience, and dataset
- Deliverables: the specific items you are required to produce for this project
- Grading: how your project will be scored
- Getting Started: guidance for how to begin working

Project Overview

For this project, you will use multiple linear regression modeling to analyze house sales in a northwestern county.

Business Problem

It is up to you to define a stakeholder and business problem appropriate to this dataset.

If you are struggling to define a stakeholder, we recommend you complete a project for a real estate agency that helps homeowners buy and/or sell homes. A business problem you could focus on for this stakeholder is the need to provide advice to homeowners about how home renovations might increase the estimated value of their homes, and by what amount.

The Data

This project uses the King County House Sales dataset, which can be found in `kc_house_data.csv` in the data folder in this assignment's GitHub repository. The description of the column names can be found in `column_names.md` in the same folder. As with most real world data sets, the column names are not perfectly described, so you'll have to do some research or use your best judgment if you have questions about what the data means.

It is up to you to decide what data from this dataset to use and how to use it. If you are feeling overwhelmed or behind, we recommend you **ignore** some or all of the following features:

- `date`
- `view`
- `sqft_above`
- `sqft_basement`
- `yr_renovated`

- `zipcode`
- `lat`
- `long`
- `sqft_living15`
- `sqft_lot15`

Key Points

- **Your goal in regression modeling is to yield findings to support relevant recommendations. Those findings should include a metric describing overall model performance as well as at least two regression model coefficients.** As you explore the data and refine your stakeholder and business problem definitions, make sure you are also thinking about how a linear regression model adds value to your analysis. "The assignment was to use linear regression" is not an acceptable answer! You can also use additional statistical techniques other than linear regression, so long as you clearly explain why you are using each technique.
- **You should demonstrate an iterative approach to modeling.** This means that you must build multiple models. Begin with a basic model, evaluate it, and then provide justification for and proceed to a new model. After you finish refining your models, you should provide 1-3 paragraphs in the notebook discussing your final model.
- **Data visualization and analysis are no longer explicit project requirements, but they are still very important.** In Phase 1, your project stopped earlier in the CRISP-DM process. Now you are going a step further, to modeling. Data visualization and analysis will help you build better models and tell a better story to your stakeholders.

Deliverables

There are three deliverables for this project:

- A **non-technical presentation**
- A **Jupyter Notebook**
- A **GitHub repository**

The deliverables requirements are almost the same as in the Phase 1 Project, and you can review those extended descriptions here. In general, everything is the same except the "Data Visualization" and "Data Analysis" requirements have been replaced by "Modeling" and "Regression Results" requirements.

Non-Technical Presentation

Recall that the non-technical presentation is a slide deck presenting your analysis to ***business stakeholders***, and should be presented live as well as submitted in PDF form on Canvas.

We recommend that you follow this structure, although the slide titles should be specific to your project:

1. Beginning
    - Overview
    - Business and Data Understanding
2. Middle
    - **Modeling**
    - **Regression Results**
3. End
    - Recommendations
    - Next Steps
    - Thank you

Make sure that your discussion of modeling and regression results is geared towards a non-technical audience! Assume that their prior knowledge of regression modeling is minimal. You don't need to explain how linear regression works, but you should explain why linear regression is useful for the problem context. Make sure you translate any metrics or coefficients into their plain language implications.

The graded elements for the non-technical presentation are the same as in Phase 1.

Jupyter Notebook

Recall that the Jupyter Notebook is a notebook that uses Python and Markdown to present your analysis to a ***data science audience***. You will submit the notebook in PDF format on Canvas as well as in `.ipynb` format in your GitHub repository.

The graded elements for the Jupyter Notebook are:

- Business Understanding
- Data Understanding
- Data Preparation
- **Modeling**
- **Regression Results**
- Code Quality

GitHub Repository

Recall that the GitHub repository is the cloud-hosted directory containing all of your project files as well as their version history.

The requirements are the same as in Phase 1, except for the required sections in the `README.md`.

For this project, the `README.md` file should contain:

- Overview
- Business and Data Understanding
  - Explain your stakeholder audience here
- **Modeling**
- **Regression Results**
- Conclusion

Just like in Phase 1, the `README.md` file should be the bridge between your non technical presentation and the Jupyter Notebook. It should not contain the code used to develop your analysis, but should provide a more in-depth explanation of your methodology and analysis than what is described in your presentation slides.

Grading

To pass this project, you must pass each project rubric objective. The project rubric objectives for Phase 2 are:

1. Attention to Detail
2. Statistical Communication
3. Data Preparation Fundamentals
4. Linear Modeling

Attention to Detail

Just like in Phase 1, this rubric objective is based on your completion of checklist items. ***In Phase 2, you need to complete 70% (7 out of 10) or more of the checklist elements in order to pass the Attention to Detail objective.***

NOTE THAT THE PASSING BAR IS HIGHER IN PHASE 2 THAN IT WAS IN PHASE 1!

The standard will increase with each Phase, until you will be required to complete all elements to pass Phase 5 (Capstone).

Exceeds Objective

80% or more of the project checklist items are complete

Meets Objective (Passing Bar)

70% of the project checklist items are complete

Approaching Objective

60% of the project checklist items are complete

Does Not Meet Objective

50% or fewer of the project checklist items are complete

Statistical Communication

Recall that communication is one of the key data science "soft skills". In Phase 2, we are specifically focused on Statistical Communication. We define Statistical Communication as:

> Communicating **results of statistical analyses** to diverse audiences via writing and live presentation

Note that this is the same as in Phase 1, except we are replacing "basic data analysis" with "statistical analyses".

High-quality Statistical Communication includes rationale, results, limitations, and recommendations:

- **Rationale:** Explaining why you are using statistical analyses rather than basic data analysis
  - For example, why are you using regression coefficients rather than just a graph?
  - What about the problem or data is suitable for this form of analysis?
  - For a data science audience, this includes your reasoning for the changes you applied while iterating between models.
- **Results:** Describing the overall model metrics and feature coefficients
  - You need at least one overall model metric (e.g. r-squared or RMSE) and at least two feature coefficients.
  - For a business audience, make sure you connect any metrics to real-world implications. You do not need to get into the details of how linear regression works.
  - For a data science audience, you don't need to explain what a metric is, but make sure you explain why you chose that particular one.
- **Limitations:** Identifying the limitations and/or uncertainty present in your analysis
  - This could include p-values/alpha values, confidence intervals, assumptions of linear regression, missing data, etc.
  - In general, this should be more in-depth for a data science audience and more surface-level for a business audience.
- **Recommendations:** Interpreting the model results and limitations in the context of the business problem
  - What should stakeholders *do* with this information?

Exceeds Objective

Communicates the rationale, results, limitations, and specific recommendations of statistical analyses

> See above for extended explanations of these terms.

Meets Objective (Passing Bar)

Successfully communicates the results of statistical analyses without any major errors

> The minimum requirement is to communicate the *results*, meaning at least one overall model metric (e.g. r-squared or RMSE) as well as at least two feature coefficients. See the Approaching Objective section for an explanation of what a "major error" means.

Approaching Objective

Communicates the results of statistical analyses with at least one major error

A major error means that some aspect of your explanation is fundamentally incorrect. For example, if a feature coefficient is negative and you say that an increase in that feature results in an increase of the target, that would be a major error. Another example would be if you say that the feature with the highest coefficient is the "most statistically significant" while ignoring the p-value. One more example would be reporting a coefficient that is not statistically significant, rather than saying "no statistically significant linear relationship was found"

"**If a coefficient's t-statistic is not significant, don't interpret it at all.** You can't be sure that the value of the corresponding parameter in the underlying regression model isn't really zero." *DeVeaux, Velleman, and Bock (2012), Stats: Data and Models, 3rd edition, pg. 801*. Check out this website for extensive additional examples of mistakes using statistics.

The easiest way to avoid making a major error is to have someone double-check your work. Reach out to peers on Slack and ask them to confirm whether your interpretation makes sense!

Does Not Meet Objective

Does not communicate the results of statistical analyses

It is not sufficient to just display the entire results summary. You need to pull out at least one overall model metric (e.g. r-squared, RMSE) and at least two feature coefficients, and explain what those numbers mean.

Data Preparation Fundamentals

We define this objective as:

Applying appropriate **preprocessing** and feature engineering steps to tabular data in preparation for statistical modeling

The two most important components of preprocessing for the Phase 2 project are:

- **Handling Missing Values:** Missing values may be present in the features you want to use, either encoded as `NaN` or as some other value such as `"?"`. Before you can build a linear regression model, make sure you identify and address any missing values using techniques such as dropping or replacing data.
- **Handling Non-Numeric Data:** A linear regression model needs all of the features to be numeric, not categorical. For this project, ***be sure to pick at least one non-numeric feature and try including it in a model.*** You can identify that a feature is currently non-numeric if the type is `object` when you run `.info()` on your dataframe. Once you have identified the non-numeric features, address them using techniques such as ordinal or one-hot (dummy) encoding.

There is no single correct way to handle either of these situations! Use your best judgement to decide what to do, and be sure to explain your rationale in the Markdown of your notebook.

Feature engineering is encouraged but not required for this project.

Exceeds Objective

Goes above and beyond with data preparation, such as feature engineering or merging in outside datasets

One example of feature engineering could be using the `date` feature to create a new feature called `season`, which represents whether the home was sold in Spring, Summer, Fall, or Winter.

One example of merging in outside datasets could be finding data based on ZIP Code, such as household income or walkability, and joining that data with the provided CSV.

Meets Objective (Passing Bar)

Successfully prepares data for modeling, including converting at least one non-numeric feature into ordinal or binary data and handling missing data as needed

As a reminder, you can identify the non-numeric features by calling `.info()` on the dataframe and looking for type `object`.

Your final model does not necessarily need to include any features that were originally non-numeric, but you need to demonstrate your ability to handle this type of data.

Approaching Objective

Prepares some data successfully, but is unable to utilize non-numeric data

If you simply subset the dataframe to only columns with type `int64` or `float64`, your model will run, but you will not pass this objective.

Does Not Meet Objective

Does not prepare data for modeling

Linear Modeling

According to Kaggle's 2020 State of Data Science and Machine Learning Survey, linear and logistic regression are the most popular machine learning algorithms, used by 83.7% of data scientists. They are small, fast models compared to some of the models you will learn later, but have limitations in the kinds of relationships they are able to learn.

In this project you are required to use linear regression as the primary statistical analysis, although you are free to use additional statistical techniques as appropriate.

Exceeds Objective

Goes above and beyond in the modeling process, such as recursive feature selection

Meets Objective (Passing Bar)

Successfully builds a baseline model as well as at least one iterated model, and correctly extracts insights from a final model without any major errors

We are looking for you to (1) create a baseline model, (2) iterate on that model, making adjustments that are supported by regression theory or by descriptive analysis of the data, and (3) select a final model and report on its metrics and coefficients

Ideally you would include written justifications for each model iteration, but at minimum the iterations must be *justifiable*

For an explanation of "major errors", see the description below

Approaching Objective

Builds multiple models with at least one major error

The number one major error to avoid is including the target as one of your features. For example, if the target is `price` you should NOT make a "price per square foot" feature, because that feature would not be available if you didn't already know the price.

Other examples of major errors include: using a target other than `price`, attempting only simple linear regression (not multiple linear regression), dropping multiple one-hot encoded columns without explaining the resulting baseline, or using a unique identifier (`id` in this dataset) as a feature.

Does Not Meet Objective

Does not build multiple linear regression models

Getting Started

Please start by reviewing the contents of this project description. If you have any questions, please ask your instructor ASAP.

Next, you will need to complete the ***Project Proposal*** which must be reviewed by your instructor before you can continue with the project.

Here are some suggestions for creating your GitHub repository:

1. Fork the Phase 2 Project Repository, clone it locally, and work in the `student.ipynb` file. Make sure to also add and commit a PDF of your presentation to your repository with a file name of `presentation.pdf`.
2. Or, create a new repository from scratch by going to github.com/new and copying the data files from the Phase 2 Project Repository into your new repository.
    - Recall that you can refer to the Phase 1 Project Template as an example structure
    - This option will result in the most professional-looking portfolio repository, but can be more complicated to use. So if you are getting stuck with this option, try forking the project repository instead

Summary

This is your first modeling project! Take what you have learned in Phase 2 to create a project with a more sophisticated analysis than you completed in Phase 1. You will build on these skills as we move into the predictive machine learning mindset in Phase 3. You've got this!