Update

We currently have all of the data we need. At this stage, we do not have any major issues holding us back from proceeding with our work on the project.

We plan to try some analysis methods which may not be directly covered in this course, such as neural nets. Thomas has experience with neural nets, and he will be able to help implement them and explain them to Justin and Victor.

Note About Kaggle

On Kaggle, people participate in discussions in which they share their current solutions, and provide advice to each other. We do not want our project to be an extension of what someone else has already come up with, and so we will refrain from reading discussion about what models and code other people are using. We will only read discussion that clarifies the problem at hand, such as what some of the variables may mean. Additionally, we will read discussion of specific models and code for other or past Kaggle projects in order to get ideas for good general practices, without copying code specific to this project.

Getting Started

None of us has experience forecasting grocery sales. Therefore, we will do some research about general trends in the retail sector, so that we have more intuition and don't have to rely solely on the code finding patterns. Justin did an internship over last summer on the strategy team of a discount grocery chain that introduced him to some of the general ideas behind complimentary and substitutionary goods as well as the effects of discounting, but nothing that will help specifically with this project.

1: Group Members

Member Name	Group Dynamic Role	Project Role
Victor de Fontnouvelle	Task Manager	Director of Research
Justin Weltz	Project Manager	Reporter
Thomas Thornton	Facilitator	Director of Computation

2: Title

Using Regression Models to Forecast Grocery Sales

3: Purpose

Forecasting sales is difficult, yet vitally important for grocery stores. Many items are perishable, and need to be sold soon after arrival. If the store orders too many supplies, they will have to discard them. If they don't order enough, they will miss out on potential sales when they run out of stock. Corporacion Favorita, an Ecuadorian grocer, currently uses very subjective methods to predict how much stock to supply their stores with. They don't analyze data. They posted a challenge on Kaggle to get help forecasting sales. We hope to be able to predict demand given a certain good, a certain store, and a certain date.

In pursuit of accurate forecasting, we will try several different methods of analysis. A secondary purpose is thus to determine which methods of analysis are best suited to this problem.

4: Data

We are provided with data from hundreds of stores and over 200,000 different products. This data is clean and ready to use.

5: Variables

Here are the data tables we are given:

sales

- date
- store - there are 54 store
- item - 4100 items
- num_sales - 83448 transactions

stores

- city
- state
- type - there are 5 types - unclear exactly what they are though
- cluster (grouping of similar stores)

items

- family - type of good - what section of the grocery store it would be placed in
- class - unclear, but there are many levels
- perishable - a binary variable

oil (has a large impact on Ecuador's current economic well-being)

- date- 1218 observations

- Oil_price

holiday_events

The test data will be given in the same format as the “sales” table. We need to predict num_sales given a specific date, store, and item. We will have to join these different data sets in order to conduct a comprehensive analysis.

6: End Product

We will try several different algorithms, and report on how each one fared, and what the optimal algorithm ended up being. Thus we will address both our main goal of forecasting sales, as well as our subgoal of determining which algorithms are best-suited to this type of problem. Algorithms we plan to try include, but are not necessarily limited to:

- Regression
- Random forest
- K-nearest-neighbors
- Neural networks

We will also format our presentation like a consulting project. This will entail making very concrete recommendations and presenting our findings in a visually appealing and understandable format.