Problem Statement

Welcome to the Wayfair Datathon! This document explains the topic of the Datathon, important details about the datasets you'll be using, and guidance on how to submit your results.

Background

In August 2011, venture capitalist Marc Andreessen published an essay titled "Why Software Is Eating The World". According to Andreessen, industries of all types would inevitably be flipped upside-down by software-driven upstarts. Expedia and Kayak replaced travel agents, Netflix replaced video entertainment, and Skype transformed telecommunication, to name just a few examples. Industry by industry, entire sectors of the economy were being reinvented.

This wave has continued to e-commerce, where entire categories of retail goods are being eaten by software companies. Amazon started with books, and then extended into basics like diapers and batteries. Meanwhile, Zappos redefined how consumers purchased shoes, while eBay captured the obscure goods market.

The furniture and homegoods market encountered similar disruption, but was fundamentally different than other retail categories. These products feature high price points and logistics challenges, while also requiring a highly-personalized consumer shopping experience. Consumers desired custom, non-branded items, that fit uniquely within their own homes.

In August 2002, Niraj Shah and Steve Conine founded Wayfair, building a platform specifically tailored to home goods shoppability. Wayfair understood that home goods e-commerce required a different buyer experience than other retail goods, and developed proprietary data science and visualization technologies specifically suited towards meeting this need.

Today, the home furniture market has exploded to over $600 billion in the U.S. and Europe alone, and Wayfair is the largest company in the space. In this growing market, Wayfair has greater annual sales and year-over-year growth than Amazon. As Wayfair nears its 20th anniversary, it will be exciting to see how it continues pushing the frontier.

Your Task

Your goal is to analyze the Wayfair e-commerce clickstream data (described in detail below), potentially in combination with supplementary datasets, in order to increase the understanding of how various factors influence customer purchasing patterns on the Wayfair online platform.

We have partially pre-cleaned several supplementary datasets for your use. Additional data is available, including details about zipcode-level real estate rental & sales prices as well as tax returns info.

You are asked to pose your own question and answer it using the available datasets in the available time. What is important is the insightfulness and depth of your conclusions and analysis. **You need not be comprehensive; quality data analysis will be rewarded over breadth of the question posed.**

Submissions may be predictive, using machine learning and/or time series analysis to predict or model online purchasing trends. Submissions may also be illuminating, through use of thoughtfully chosen data visualizations or sound statistical tests.

Consider exploring one of the sample questions below, or creating your own variation. Creativity in formulating your own question generally has a positive effect on judges' assessment of your submission; **however, it should not be at the expense of analytical depth, precision, and rigor, which are far more important.**

Sample Question 1: What characteristics are most predictive of the likelihood to buy something online from Wayfair?

Sample Question 2: Explore the online clickstream journey of buyers and non-buyers. Are there any interesting patterns of note?

Sample Question 3: What sorts of relationships exist between the demographics of certain zip codes and the amount and types of items that are bought?

Sample Question 4: Does Wayfair's selling performance differ notably in areas where there are competitor products? What product niches does Wayfair do best against the competition?

Datasets

The provided datasets are stored in the "Datathon Materials" folder on Box and are spread across eight tables. Your team should only use the tables that are relevant to your chosen question/topic. The raw data sources are noted; however, we encourage you to use our tables since they have been organized and cleaned to "play nice" with each other.

clickstreams_with_purchase

Information about the series of pages that users clicked on which resulted in a purchase on the Wayfair platform. *Note: in order to keep the dataset size manageable, the provided data is a*

25% unbiased sample of the raw data. If using click count metrics, remember to multiply quantities by 4 to approximate the actual data.
~11 million rows & 6 columns. Size: ~85MB zipped, ~850MB unzipped. Source: Wayfair.

clickstreams_without_purchase
Information about the series of pages that users clicked on which did not result in a purchase on the Wayfair platform. *Note: in order to keep the dataset size manageable, the provided data is a 1.25% unbiased sample of the raw data. If using click count metrics, remember to multiply quantities by 80 to approximate the actual data.*
~11 million rows & 5 columns. Size: ~90MB zipped, ~850MB unzipped. Source: Wayfair.

orders
Information about Wayfair online orders during a single week in July 2018.
~140,000 rows & 9 columns. Size: ~12MB. Source: Wayfair.

products
Information about Wayfair products sold online.
~500,000 rows & 15 columns. Size: . Source: Wayfair.

comp_products
Information about products of Wayfair's competitors that are sold online.
~1.25 million rows & 6 columns. Size: ~200MB zipped, ~500MB unzipped. Source: Wayfair.

rental_prices
Data about the rental prices of various property types across U.S. zip codes in 2018.
19,782 rows & 17 columns. Size: ~2MB. Source: Zillow.

sale_prices
Data about the sale prices of various property types across U.S. zip codes in 2018.
42,735 rows & 17 columns. Size: ~4MB. Source: Zillow.

taxes
Data containing details about the tax returns filed across U.S. zip codes in 2016.
~180,000 rows & 54 columns. Size: ~30MB. Source: Internal Revenue Service.

Additional Datasets

You are welcome to scour the Web for custom datasets to supplement your analysis. All additional data used should be public and should not exceed 2GB unzipped (consult Correlation One's technical product team if you believe your idea is worthy of an exception).

Other Materials

We will provide you the schema for each of the data tables in another packet.

We will also provide you a Datathon manual at registration, which contains a section on using Box. This will show you how to download the datasets (described above) and upload your submissions (described below).

Submissions: Content

Submissions should have two components:

1. Report – this should have two main sections:
    a. Non-Technical Executive Summary – What is the question that your team set out to answer? What were your key findings, and what is their significance? You must communicate your insights clearly – summary statistics and visualizations are encouraged if they help explain your thoughts.
    b. Technical Exposition – What was your methodology/approach towards answering the questions? Describe your data exploration process, as well as your analytical and modeling steps. Again, use of visualizations is highly encouraged when appropriate.
2. Code – please include all relevant code that was used to generate your results. **Although your code will not be graded, you MUST include it or your entire submission will be discarded.**

Additional information (e.g. roadblocks encountered, caveats, future research areas, and unsuccessful analysis pathways) may be placed in an appendix.

Judges will be evaluating your work without your team there to explain it; therefore, **your submission must "speak for itself"**. It need not be polished to the level of a final product, but do ensure that your main findings are clear and that any visualizations are functionally labeled.

Submissions: Evaluation

You will be evaluated based on your Report, as follows:

- **Non-Technical Executive Summary**
    - *Insightfulness of Conclusions.* What is the question that your team set out to answer, and how did you choose it? Are your conclusions precise and

nuanced, as opposed to blanket (over)generalizations?

- **Technical Exposition**
  - *Wrangling & Engineering Process.* Did you conduct proper quality control and handle common error types? What sorts of feature engineering did you perform? Please describe your process within your Report. Note that we care about the structure of your data engineering process and the choices you made, **NOT** detailed mechanics or the specific code you used.
  - *Investigative Depth.* How did you conduct your exploratory data analysis (EDA) process? What other hypotheses tests and ad-hoc studies did you perform, and how did you interpret the results of these? What patterns did you notice, and how did you use these to make subsequent decisions?
  - *Analytical & Modeling Rigor.* What assumptions and choices did you make, and what was your justification for them? How did you perform feature selection? If you built models, how did you analyze their performance, and what shortcomings do they exhibit? If you constructed visualizations and/or conducted statistical tests, what was the motivation behind the particular ones you built, and what do they tell you?

In lieu of the above, we recommend that your team not try to learn new tools if possible; instead, leverage your existing skills to extract as much insight from the data as you can. **Historically, the correlation between the complexity of modeling techniques used and whether a team wins one of the top prizes has been at best weakly positive - the soundness of the overall process is far more important.**

Submissions: Format

Reports can be produced using any tool you prefer (Python Notebook, Shiny Application, Microsoft Office, etc.); however, **your report MUST be in a universally accessible and readable format (HTML, PDF, PPT, Web link)**. It must not require dedicated software to open. For example, if your report is a Python Notebook, it should be exported to HTML. If you create a Shiny App, it should be published at an accessible Web link.

However, please also include the source file used to generate your report. For example, if you submit a PDF with math-type, equations, or symbols, please include your raw LaTeX source file.

Code should be submitted in a single zipped collection of files separate from your report.

Your team will be provided a sheet with your team's Box account login details when the hacking session begins; you will be using the account to download the datasets as well as to upload your submission content. We recommend that you wrap up your work by 3:15 PM and begin

uploading your submission at that time. **Submissions MUST be received by 3:30 PM. Any submission received after 3:30 PM will NOT be evaluated by the judges**.

Tips & Recommendations

You will have ~12 hours total to work on the problem statement. However, you will not have access to the full dataset until the morning of the competition. As such, we recommend you split your time as follows:

- Friday evening, ~7:00PM – 12:00AM: You will receive a copy of the problem statement, data table schema, and data table heads. This gives you the opportunity to study the available data fields, think about suitable questions to tackle, and plan out your exploration process. Additionally, the data table heads should be sufficient for you to begin putting together some data wrangling & cleaning scripts.
- Saturday, 8:30AM – 3:30PM: You will receive the actual data. If you set up your data munging scripts already, you should be able to quickly apply them and immediately begin working with the data. You should spend most of your day investigating the data, performing qualitative & quantitative analysis, and writing up your process & results.

For data engineering, exploration, and modeling, we highly recommend that you install Jupyter Notebook: http://jupyter.org/install.html. Jupyter Notebook is an interactive, real-time development environment that eliminates many pain points of the standard "terminal + text editor" environment, and is compatible with both Python and R.

Finally, **we STRONGLY encourage you to start typing up your final submission AT LEAST three hours before the submission deadline**. In the past, many teams have spent a lot of time conducting great analyses, only to realize that they left almost no time for actually writing up and presenting their results. **This cannot be stressed enough – quality data analysis that is incomplete or poorly presented will NOT win one of the top prizes**.

Ask for Help

The Datathon team is here to help. Let us know about your struggles as early on as you can and we may be able to offer advice on how to best move your analysis forward.