Section 1: Stock Market Data

1. Why might the market shift from momentum being dominant to mean reversion being dominant (or the other way around)?

Momentum assumes that there exists a trend in the market and the market will continue to follow the trend before reversing. The start of a new market trend often coincides with major market-level changes such as interest rate changes, economic activity changes, as well as the implementation of new government policy such as quantitative easing. On the other hand, the start of a new trend in a stock return may be triggered by both broad factors such as industry rotation or demographic changes as well as asset-level factors such as new product releases or the entry of new competitors.

Investors initially under-react to those changes of fundamentals because of anchoring bias. However, some smart investors act on those changes first and establish the initial trend. Then other investors gradually embrace the smart investors' view and the stock price trends. Finally, many speculators jump in and the herding effect pushes the stock to become overbought/oversold, which is usually followed by mean reversion.

Mean reversion assumes an oversold stock is likely to bounce back from its lows and an overbought stock is likely to fall from its highs. It usually happens when a trend is likely to reverse or when it becomes exhausted or overextended. Therefore, when stock prices become overbought/oversold or the existing trend becomes exhausted, the market will shift from momentum being dominant to mean reversion being dominant.

There are also two psychological biases from behavioral finance perceptive that help to explain the shifting. Investors tend to be conservative; they under-react to information because they stick to their prior beliefs. Such conservatism causes momentum. On the other hand, investors assume commonality between similar objects. Such representativeness leads to mean reversion.

2. Why might trading momentum or mean reversion succeed as a strategy? Why might it fail?

The general principle of momentum strategies is to buy high and sell higher; while the general principle of mean reversion strategies is to buy low and sell high. Mean reversion assumes the existence of a high probability attraction range, which is indicated by fundamentals. A price reversion inside this range discloses trading opportunities. Momentum assumes market inefficiency, non-stationary behavior of returns and long-term memory existence. The following narrates two factors that help explain why such seemingly opposite principles might both work in practice.

One observation from historical stock price movements is that sometimes the market shows strong trends and sometimes it trades sideways with no clear trend. If someone can identify whether the stock price will be in a trending regime or a mean-reversion regime, he can successfully benefit from both momentum strategies and mean-reversion strategies. When a momentum trade enters the market, the trader has a positive profit expectation as long as the momentum signal sufficiently increases the odds that the stock price will continue the trend after the entry. Mean reversion traders, on the other hand, try to capture the reversion when the market is likely to enter a mean-reversion regime and avoid betting against the market when the market continues the existing trend.

Another observation is that stock prices tend to exhibit long-term momentum and short-term mean-reversion patterns. Strategies that buy (sell) at short-term oversold (overbought) conditions when the long-term trend is up (down) tend to generate a high Sharpe ratio. Short-term mean-reversion signals indicate temporary overbought or oversold conditions. By trading in the direction of the longer-term momentum, investors reduce the likelihood of catching a falling knife and increase the odds that the overbought/oversold conditions will revert soon.

The risk associated with both momentum and mean-reversion strategies is to identify the appropriate timing to enter/exit the market. When new fundamentals trigger a new trend, momentum traders do not respond immediately. Instead, they wait for market confirmation of the new trend before entering the market. Similarly, mean-reversion traders wait for market confirmation of the trend reversal before exiting the market.

Section 2: Oklahoma State Spending

5. Please describe the process you would follow to build a model on this dataset to make predictions about the stock market. Please note this is a hypothetical only - there is no need to build an actual model.

Credit card transactions are usually utilized by asset managers to predict the performance of major retailers and make investment decisions. The majority of a retailer's revenue comes from its sales, which can be estimated by its customers' credit card transactions. Traditional investors rely on a company's financial report for its fundamentals such as revenue, long-term liabilities to assess its performance. However, a company's financial report is released every quarter. In order to access a retailer company's fundamentals in a timely manner, investors can utilize credit card transactions history of its customers to evaluate its fundamentals before its financial report is released. Such advanced information will help investors to predict the stock market in a more efficient way.

Given this dataset that comprises credit card transactions by employees of the State of Oklahoma, we can build a model to predict the stock price movements for local companies headquartered in the State of Oklahoma. Based on the model's prediction, a trading strategy can also be constructed. The following separately describes the general procedure of model development.

A. Data collection:

A supervised learning model needs to be developed which uses historical information from credit card transactions to predict future stock price movements. With such a goal in mind, we need to define the labels for our supervised learning model. First, we will decide which stocks we want to predict. Given the existing biases of this dataset, we may focus on local company's stocks in the State of Oklahoma. Then, we collect the historical prices of those stocks corresponding to the timeframe of the credit card transactions' history. The features (or explanatory variables) are all credit card transaction information for one day and the label will be defined based on the market-adjusted return over the next 10 trading days (1 if positive return and -1 if negative return). The whole dataset will be split into training, validation and testing datasets.

B. Data preprocessing:

This step takes the most of the time since it involves data cleaning, data manipulation, and data understanding. Since this dataset contains text data such as transaction descriptions and merchant categories, we need to clean the data and extract the important information

which can be done through NLTK and Word2Vec in Python. Since we have one-year transactions but a few columns in this dataset, we need to conduct feature engineering and add more features to improve our model's predictability. For example, customers' sentiment on some product can be reflected through the total number of purchases/returns. An average transaction volume for some merchant category can be used as a benchmark to determine one company's competitiveness. A data visualization step will also be helpful for us to detect any trends or seasonality in the transactions.

C. Model development:

Random Forest and Gradient Boosting will be considered as benchmark models because they are both ensemble methods that can handle a dataset with both numeric and text features. However, both methods have several hyper-parameters to tune. With the benchmark models aside, we will try to train this dataset into an LSTM model due to its advantage in training sequential data with built-in memory. The basic rationale is that there exists a momentum in transaction volume because of the herding effect of customers.

D. Model evaluation:

The model will predict a signed confidence value between -1 and 1. If we expect a stock price significantly raise compared to the broad market over the next ten days, the model might assign it a large, positive confidence value (near 1.0). If we expect a stock price significantly declines, the model might assign it a large, negative confidence value (near -1.0). If unsure, the model might assign it a value near zero. A synthetic portfolio can be constructed by buying stocks with positive confidence values and selling stocks with negative confidence values. We will evaluate our model's performance by the Sharpe ratio of our synthetic portfolio constructed based on the trading signals generated from the model.

6. What biases might this dataset have if you tried to use it to model equities?

This dataset comprises credit card transactions by employees of the State of Oklahoma. As we observed from the dataset, most of the customers are agencies of the State of Oklahoma and their credit card transactions are mainly based on business purposes. In other words, this dataset has a bias on the credit card transactions of the state's agencies for business use, which cannot represent the total credit card transactions made by all the residences in the State of Oklahoma.

On the other hand, this dataset contains major retailers such as Walmart which has customers in the other states or oven international customers. However, this dataset only represents the credit card transactions made in the State of Oklahoma. If we want to build a model to predict Walmart's stock price movements, the total amount spent at Walmart based on this dataset is far from a good representative of its revenue.

Moreover, when we looked at the timeframe associated with this dataset we noticed that it only contains a one-year history of credit card transactions. When evaluating the performance of a company, we don't have sufficient historical credit card transactions to compare which might lead to noisy or biased predictions.

Finally, credit card transaction might be a biased estimate of a company's fundamentals such as revenue. There are other types of transactions as well. In addition, when considering the delinquencies of customers, credit card transactions are not fully guaranteed to be part of a company's revenue. Even though such a bias might be insubstantial given the customers are mostly the state's agencies, it is not nonexistent.

7. (Optional) Do you have any other observations about this dataset?

It is interesting to observe that when breaking out the transactions into 12 months over a year, December has the least transaction volume and also the least transaction amount. One possible explanation would be December is a holiday season when employees of the State of Oklahoma take vacations and so their transactions of business use would be less than the other months.

The most merchants purchased by the employees are equipment, laboratory products, hardware, industrial supplies, etc. It is not surprising because most of the employees are working at universities, Grand River Dam Authorization, and other states agencies.

The vendor who involves the highest transaction amount is American Airline, considering the price of a flight is generally one of the highest among all the merchants. We can probably infer from this observation that a majority of employees in the State of Oklahoma take business trips with American Airlines.

Section 3: Feedback

3. (Optional) At Two Sigma, we find patterns, meaning and relationships in the world's data to create value for our investors. We are constantly on the lookout for new and interesting datasets. If you know of a dataset that might be of interest to Two Sigma, please describe it here (with any relevant links).

News Data: RavenPack scans news content from a variety of sources (newswires, corporate releases, papers, rating agencies, investment banks, and regulatory agencies), processes it in real time using natural language processing algorithms and publishes the extracted information. Web scraping is a common approach to extracting news from websites.

Social Media Data: Google Trends provides weekly and daily updates on the search volumes of different search terms such as industrial activities, company names, and brands. Therefore, Google Trends can provide timely information on economic activities and consumer interests for specific industries and companies.

Geolocation Data: FourSquare provides geolocation data that tracks the foot traffic of customers. Based on this information, investors can assess the performance of a retailer store or a restaurant by analyzing the patterns of its customers' foot traffic.

Satellite Data: Orbital Insight track the ongoing health of major retailers by analyzing satellite images of their parking lots. The idea is that a greater number of cars per week or month observed in the parking lots of a retailer indicates strong customer interest. Satellite data can also be useful for detecting real-time levels of oil inventories, crop output, and mine production, all of which could have a profound impact on trading decisions.