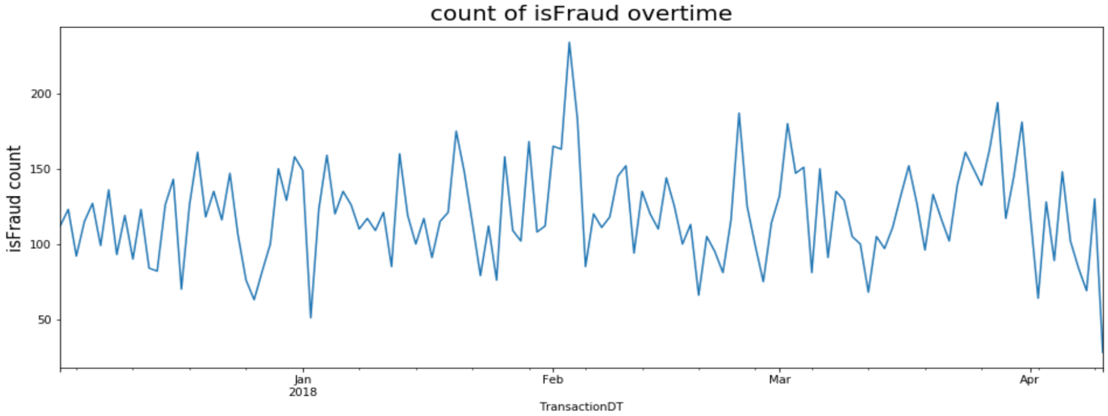
Capstone Project 1 – Exploratory Data Analysis

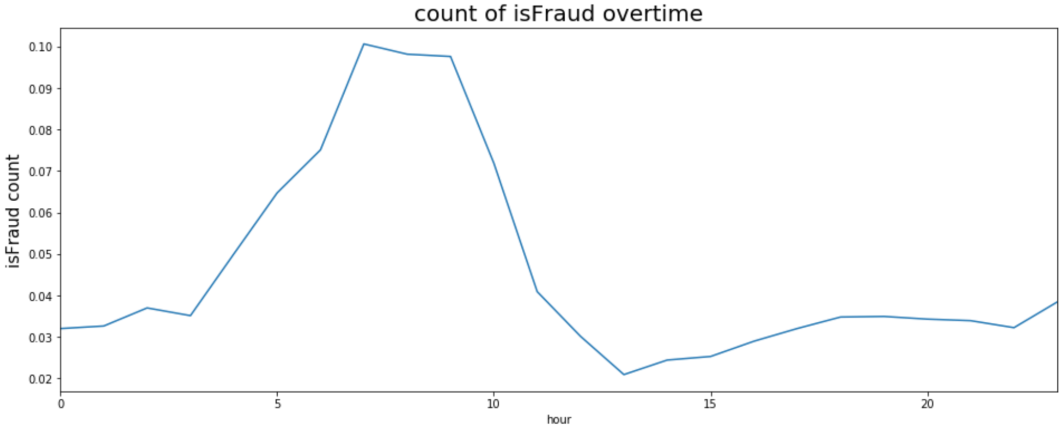
Fraud detection is significant technique in a financial service company. This capstone project is a Kaggle closed competition which focus on how to improve the accuracy of credit card fraud detection. From the data sets, there are three questions to follow up by analyzing the credit card transaction and identity data.

1. Are there variables that are particularly significant in terms of explaining the answer to your project question?

   I've noticed there are features about the date and time in the credit card transaction data set. I want to measure if the fraud action is somehow related to the time of transaction. Therefore, I draw a fraud count graph with daytime and try to find if there's anything special:

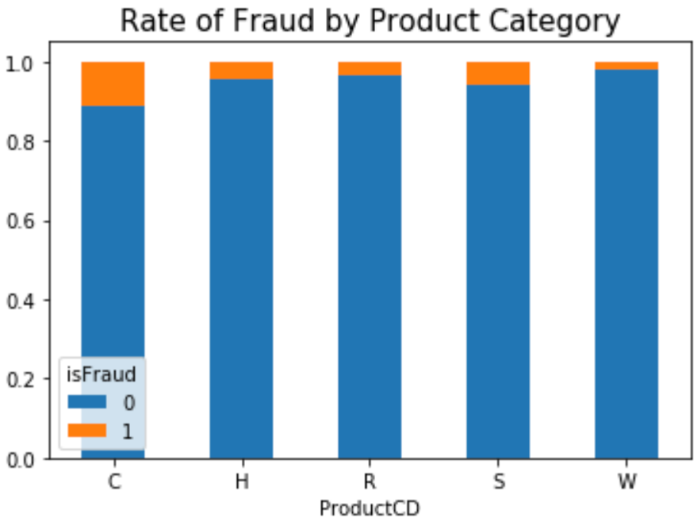


   According to the figure above: no significant pattern visually based on days. Thus, I changed the unit to hours:



Then we can notice that there's a sharp increase between 5 am to 10 am. Thus we can conclude that the fraud action seems to peak during early morning.

In addition, I've noticed that different product type may cause more or less frauds as the following figure shows:



2. Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?

   In the transaction information table, there are 6 card information such as card type, card category, issued bank, or country. I think there're strong correlations between some of the card features. Therefore, we can use Pearson coefficient or scatter plot to determine whether they have strong correlations between each other.

3. What are the most appropriate tests to use to analyze these relationships?

   I think the most appropriate test for this problem is Hypothesis Test. We can set there are no strong correlations between this 6-card information as null hypothesis, and they're strongly correlated to each other as alternative hypothesis. Then use Pearson coefficient to run the test.