Northeastern University
D'Amore-McKim
School of Business

Professor Kate Ashley
MISM 6202

Foundations of Data Analysis for Business
PROBLEM SET 1

Use the file 'commutervan.csv' for all questions. Refer to 'commutervan_data_dictionary.csv' for descriptions of the variables included in this dataset.

Commuter Van Express, Inc. (CVE) is a commuter shuttle service with operations in a large US city where public transportation is frequently delayed and overcrowded. The company operates a fleet of 14-passenger vans, equipped with WiFi and comfortable seats, to provide shuttle service along fixed routes between residential neighborhoods and the city center during commuting hours (running toward the city center in the morning and towards the residential areas in the evening). CVE's customers use a website or mobile app to book a seat in one of the shuttles, thereby guaranteeing that space will be available. Vans run on a regular schedule along each route, and the app includes location tracking to provide users with real-time arrival information.

CVE uses two different analytics platforms to collect and organize data on their ongoing operations. Platform 1 tracks ride volume and revenues, which can vary per ride due to volume discounts on package purchases (e.g. 12 rides for the price of 10), a monthly subscription option, and various short-term coupons and promotions. Platform 2 tracks user activity in the mobile app, including actions like starting a new session, tapping on a stop, booking a ride, etc.

It is April 1, 2016 and CVE has hired you as a consultant to help them understand their recent performance and develop a method to forecast future rides and revenues. To assist in your analysis, the company has provided you with daily data from both of its analytics platforms for the first quarter of 2016. The data appear to be clean and free of errors, with the exception of two days for which the app activity metrics are missing (you've been told that there was app activity on those days, but usage statistics were not recorded due to a system error).

Use RStudio to answer the following questions. Provide your written answers, along with any relevant tables and charts, in a **single PDF file**. Any charts included in your report should be properly labeled and formatted for an audience of company executives. You should also submit a **single .R script file** with your code for the analysis.

Regression Analysis.

1. Because customers value flexibility in their commuting plans, CVE allows customers to cancel a booking without penalty up until the van they booked arrives at their chosen stop. As a result, not all ride bookings result in a ride actually taking place. Estimate a simple linear regression model to understand the relationship between daily bookings and daily completed rides. Report the estimated regression equation and R^2 value and interpret them in words.

2. Estimate a simple linear regression model to understand the relationship between daily completed rides and daily revenues. Report the estimated regression equation and R^2 value and interpret them in words.

3. CVE would like to know if ride bookings through the mobile app can be predicted using the actions that an app user may perform prior to booking: namely, starting a session, tapping on a stop, tapping on the sidebar, and viewing van ETAs. Estimate a multiple regression model that uses the relevant variables to predict ride bookings. Multiple models involving these variables are possible; select the best model and explain your choice, citing specific numerical evidence from the regression output. Report the estimated regression equation and R^2 value and interpret them in words.

Forecasting.

4. Construct a *k*-period simple moving average that can be used to forecast the number of daily completed rides, where *k* is chosen based on your assessment of the seasonality patterns in the data. Explain your choice of *k* and report MSE, MAD, and MAPE for this forecasting model.

5. Construct a simple exponentially smoothed series to forecast number of daily completed rides, using initial value equal to y_1 (the first value of 'rides' in the dataset) and your choice of smoothing parameter α. Explain your choice of α and report MSE, MAD, and MAPE for this forecasting model.

6. Create a scatter plot to visually inspect the 'rides' variable. Describe any trend and seasonality that appear to be present.

7. Estimate a linear trend model for number of daily completed rides. Report the estimated linear trend equation and the R^2 of the model, and interpret both the equation and the R^2 in words.

8. Estimate and interpret a linear trend model with relevant seasonal dummy variables for number of daily completed rides. Report the adjusted R^2 and comment on its magnitude relative to the adjusted R^2 from the regression you performed in (7).

9. Use the estimated regression equation from (8) to calculate a forecast of daily completed rides for each day in your dataset. Calculate MSE, MAD, and MAPE for this forecast. Comment on which of the three forecasts you have calculated in this problem set performs the best (i.e., has the lowest forecast error).

10. Use the estimated regression equation from (8) to forecast daily completed rides for each weekday in April 2016. *Optional: Also forecast revenues for each day.*