**Project Log Sheet**

**Student’s Name:** .....Linni Qin........................ **Student’s ID:** .....n9632981

**Student’s Name:** .....Linni Qin.......................................................... **Date:** …18/08/2017………..**Meeting No: 1**…

**Project title:** IT MASTERS (DS): Perform a Data Science Analysis of a Dataset/Task **UNIT:** ……IFN701………….

**Project title:** IT MASTERS (DS): Perform a Data Science Analysis of a Dataset/Task **UNIT:** ……IFN701………….

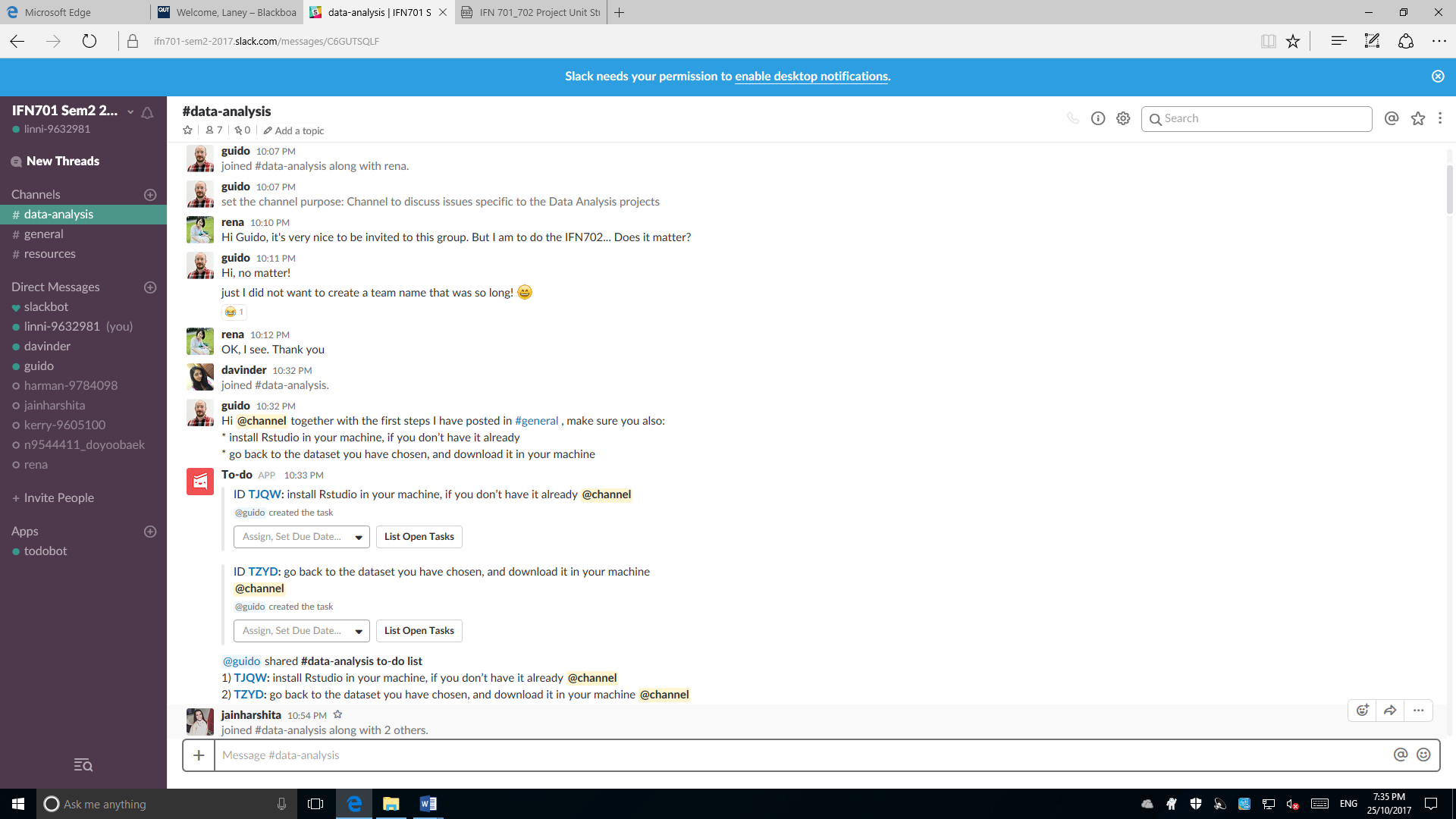
**Supervisor’s Name:** ……… Guido Zuccon …………**Signature:** ………………………………….

**I hereby to state that the following nine supporting evidences for the project log are true.**

As stated at the attached project log sheets, there are only three formal meetings with Dr. Guido Zuccon who is my supervisor of unit 701 this semester. The less chance of meeting is caused by his busy conference schedule and his sick leave. The following nine important Slack chatting screenshots are used to prove that the consistent communications focusing on my project with the supervisor.

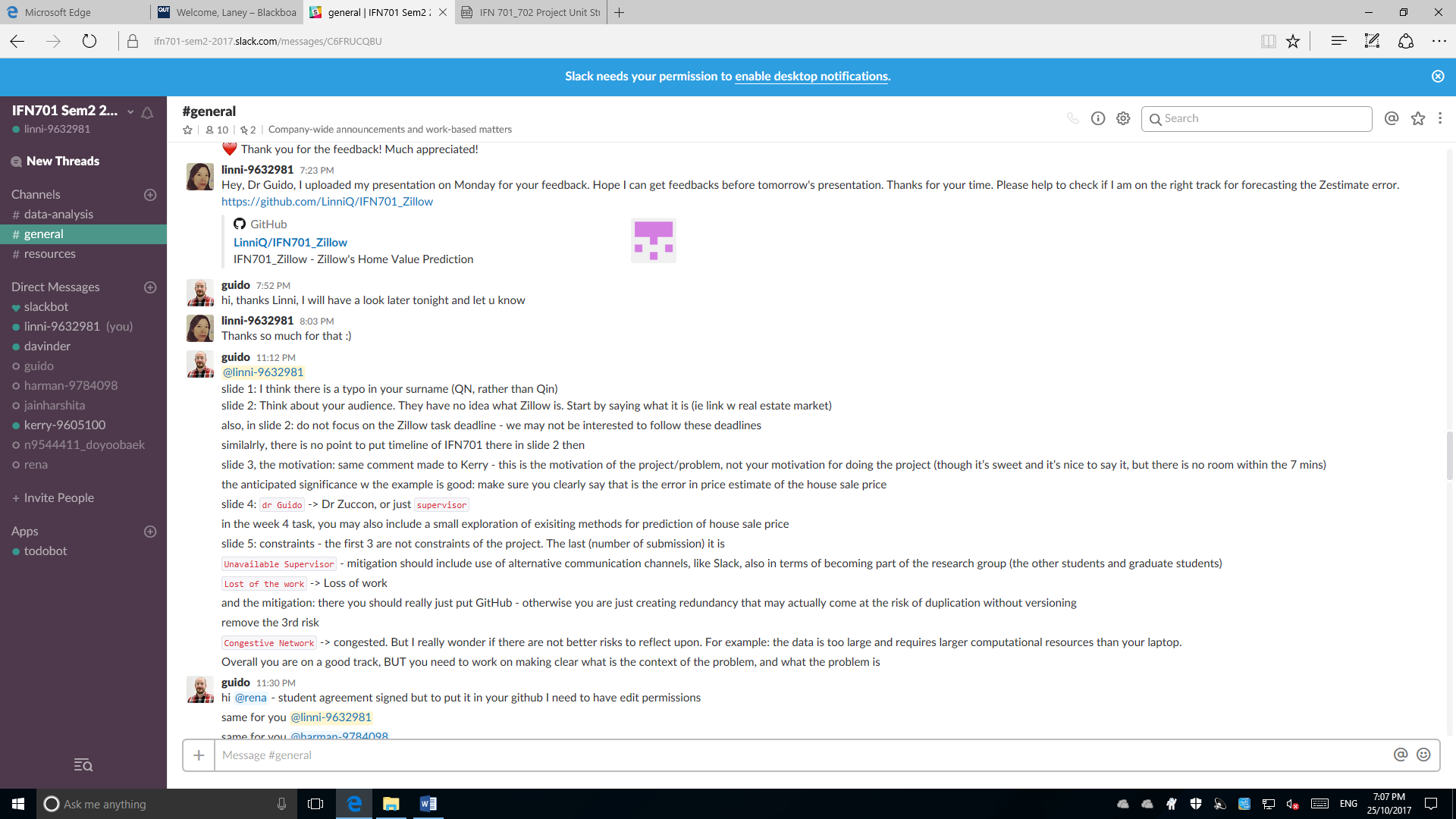
July 31, 2017 Project start-up “To do” List:

1. Prepare the R and RStudio on the laptop
2. Get the data ready
3. Get a GitHub account to save the works for supervisor’s convenience to provide the feedback
4. Introduce individual project in the team



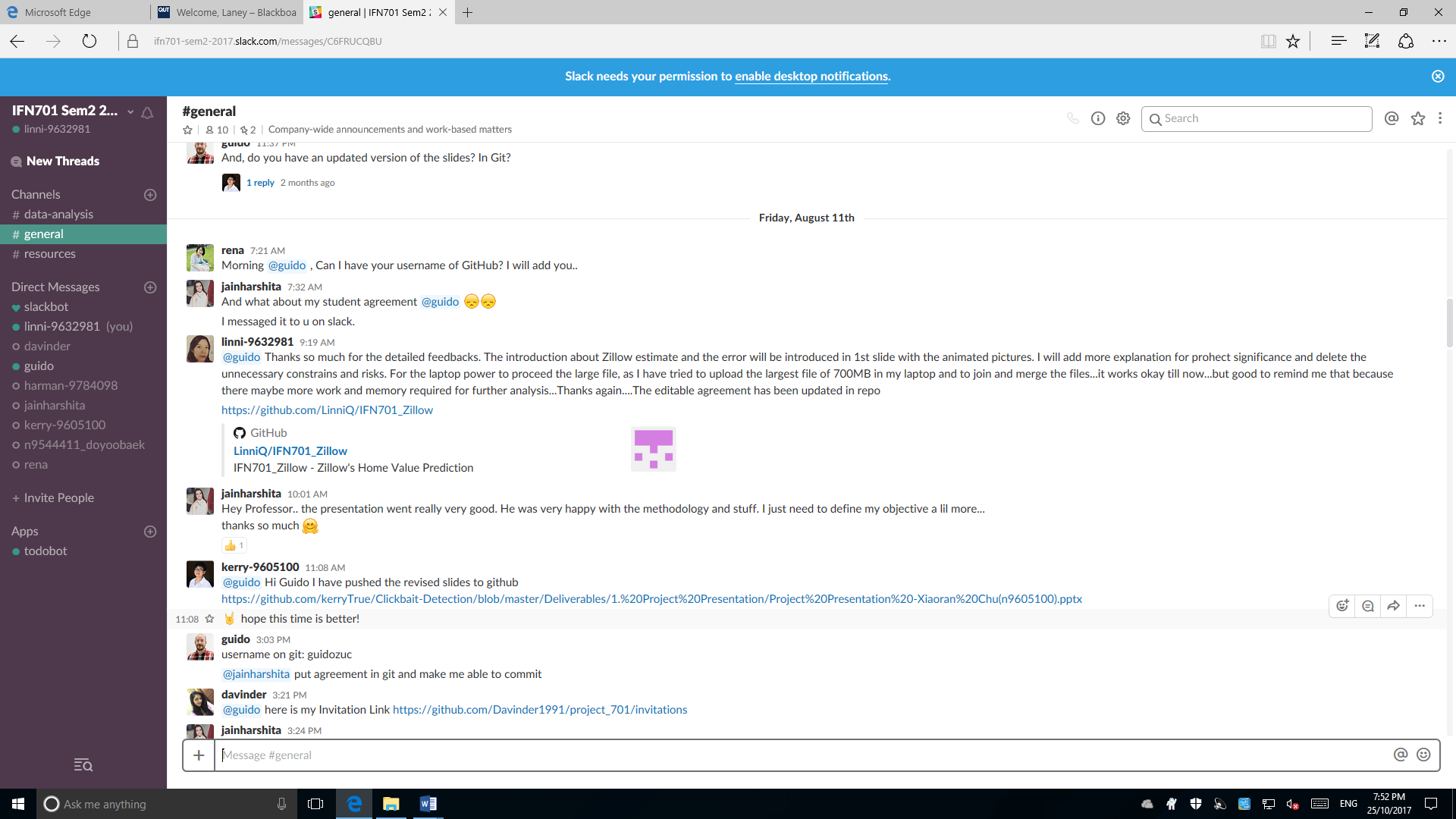
August 10, 2017 Proposal Presentation Discussion

1. Correct the typo
2. Redo the motivation of the project



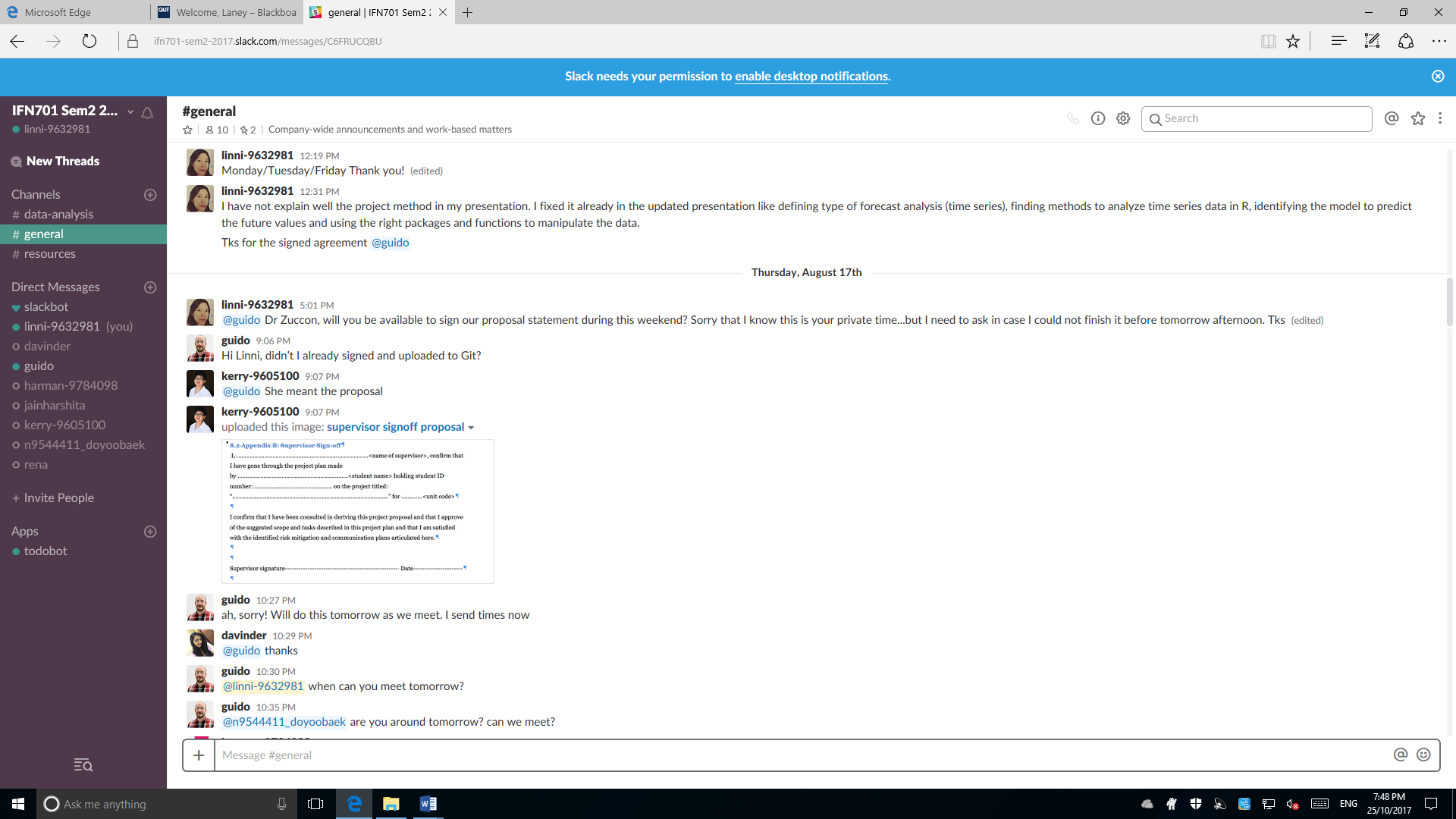
August 11, 2017 Discussion on Project Agreement

1. Update the agreement on GitHub for supervisor’s signature



August 17, 2017 Discussion on project plan

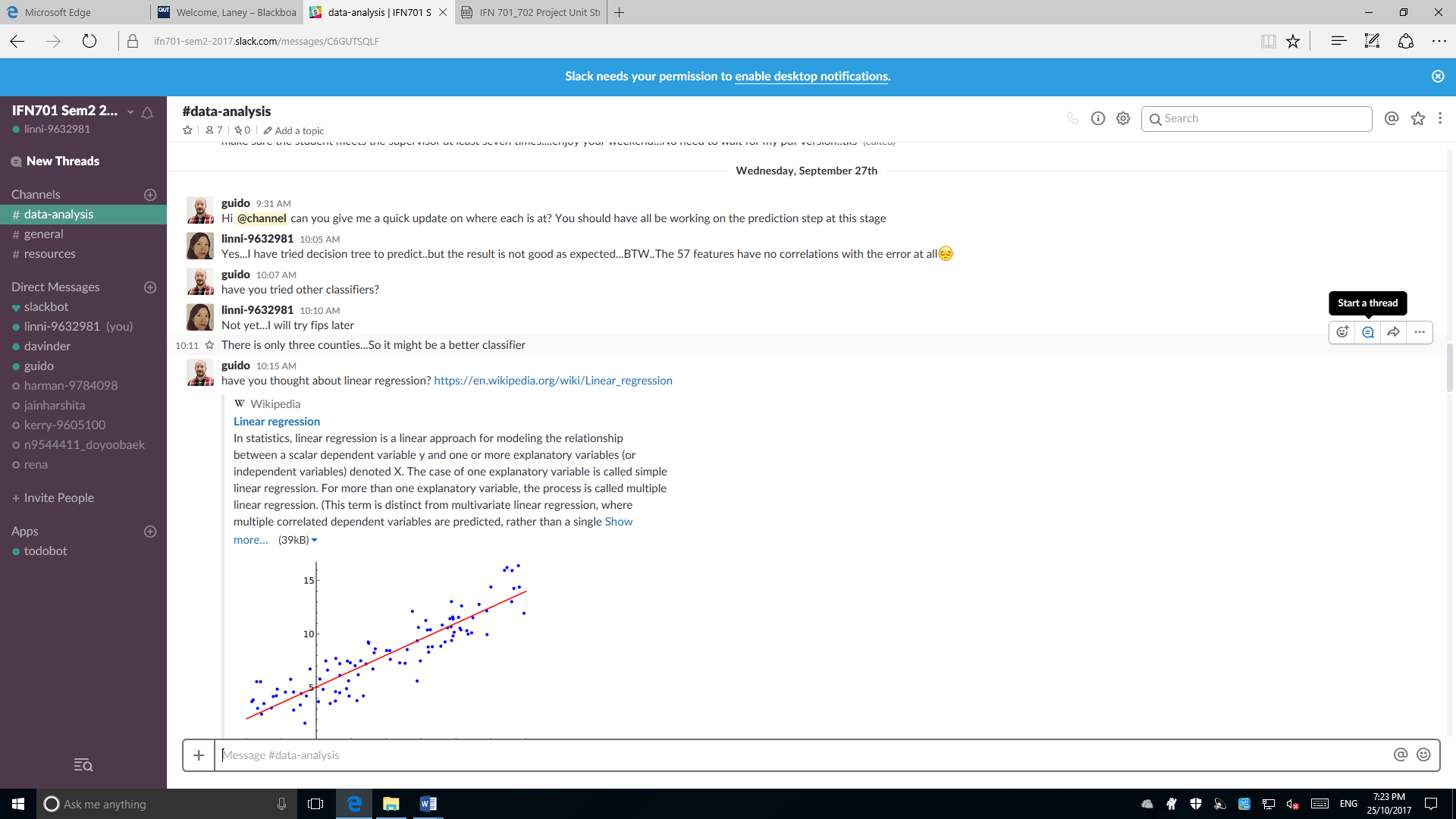
1. Get the draft ready and plan the meeting in next day



September 27, 2017

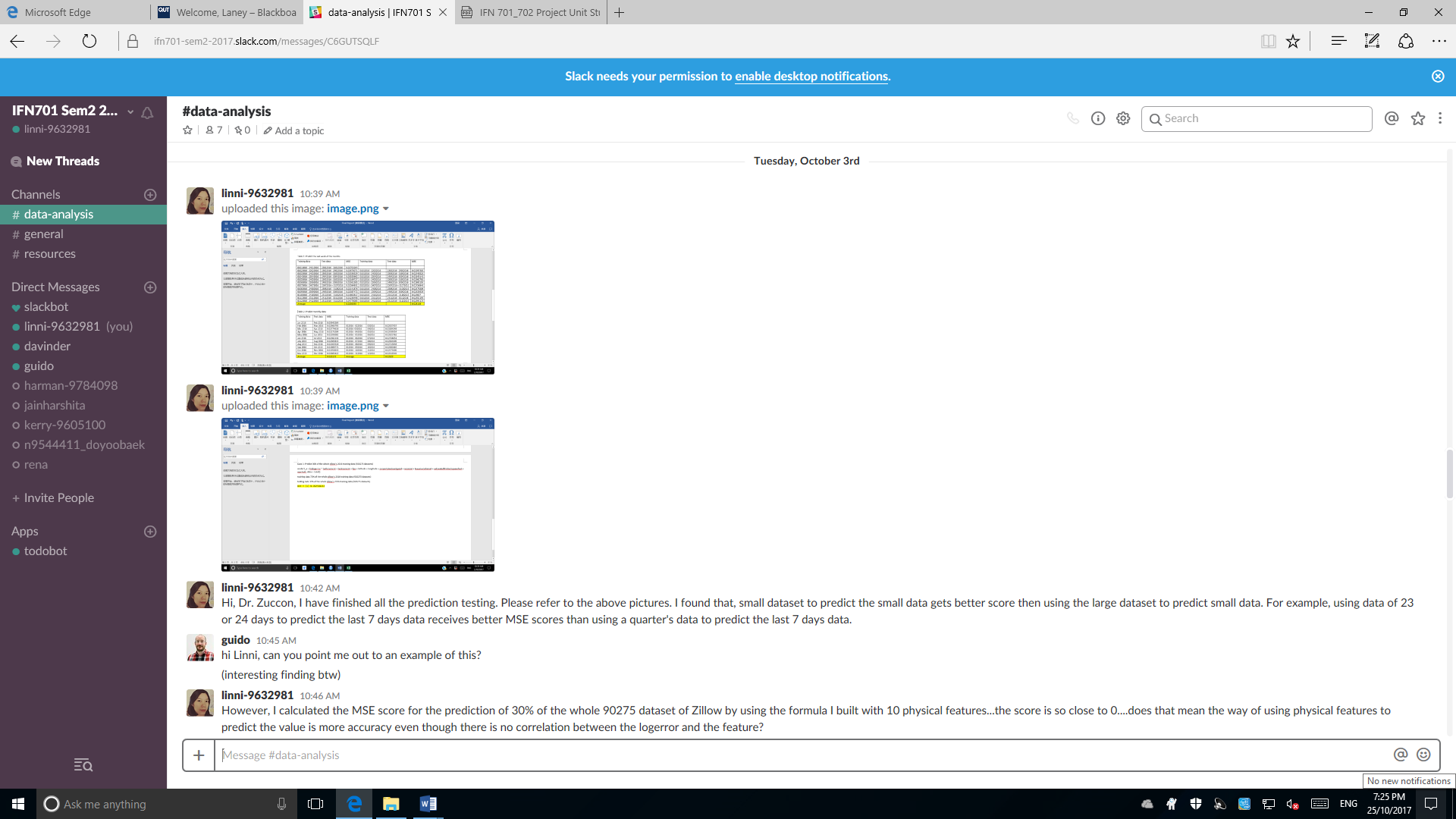
Discussion on the prediction model

1. Binary prediction models have been proved not suitable for this project
2. Linear Regression is better for this project
3. I target at the multiple linear regression model



October 3, 2017

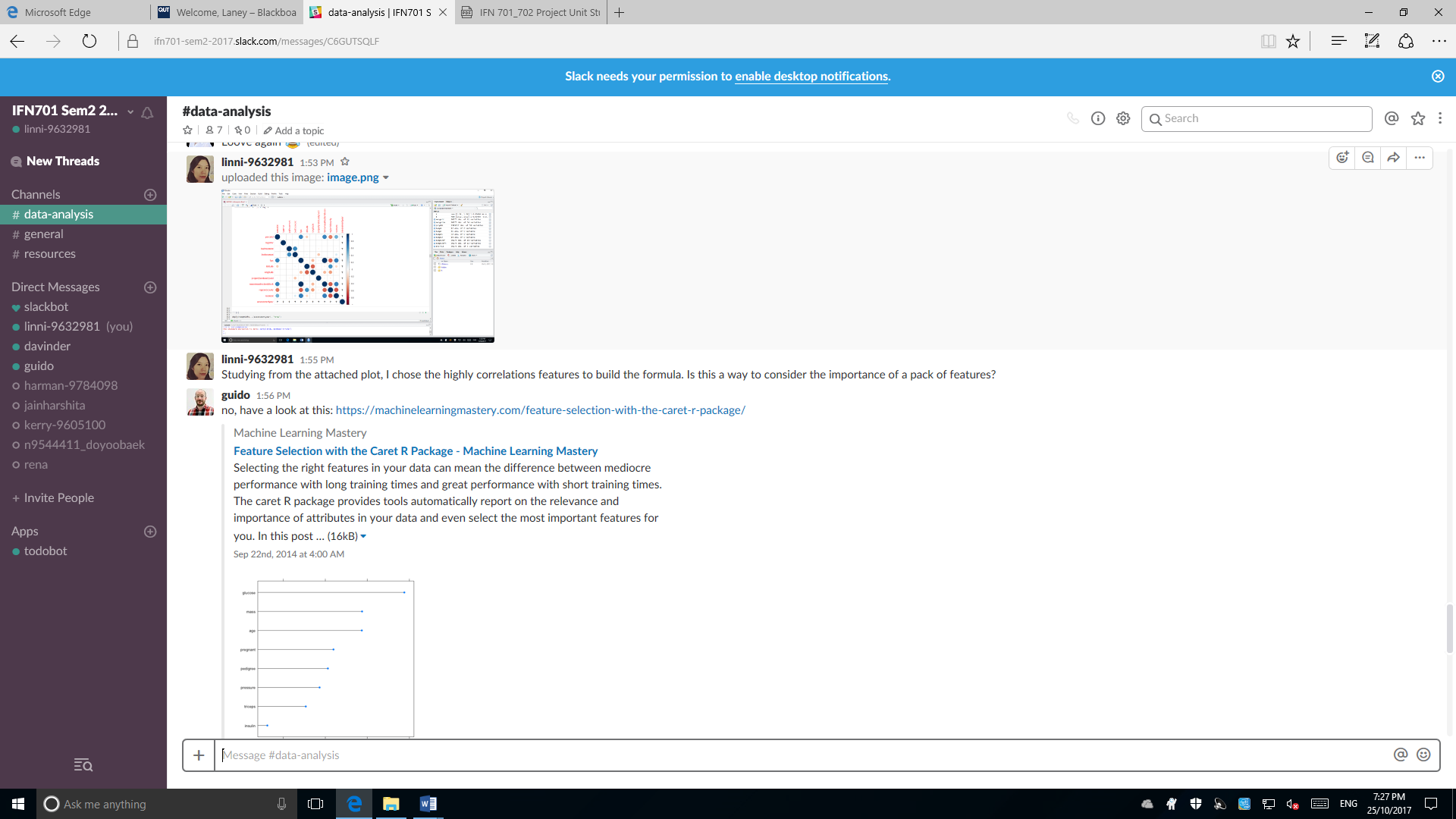
Update of the Mean Squared Error Value based on the testing on different groups of training and test datasets.



October 5, 2017

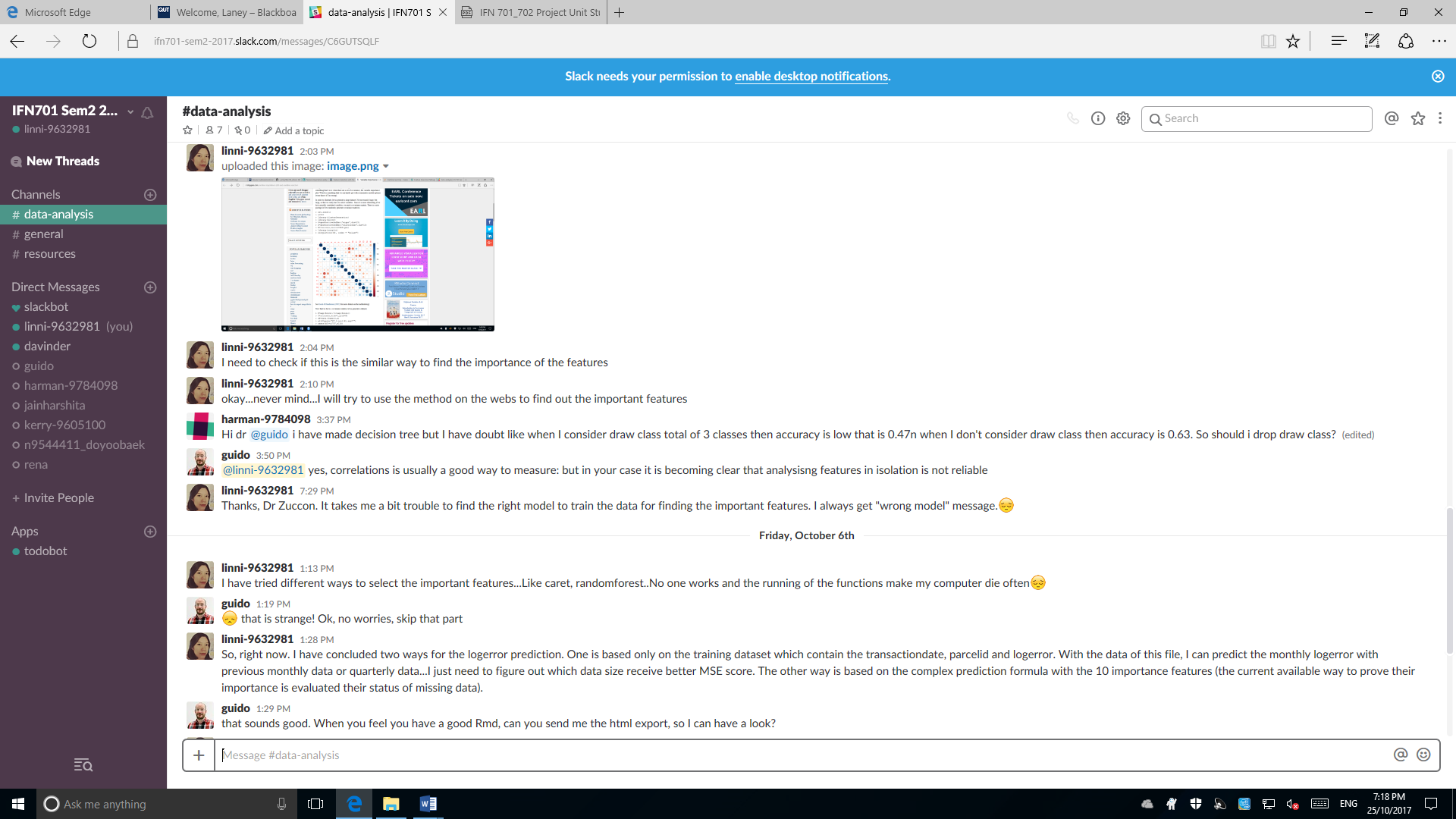
Discussion on the challenge of zero correlation between the log error and the property features.

Discussion on the obstacle of applying the package “caret” and “randomForest” to identify the importance of the variables.



October 6, 2017

Updated the results based the discussion on October 5. Supervisor approved to skip some work.



October 7 – October 13, 2017

As the new data of Zillow in 2017 was updated on October 3, I need to do more work on exploring the data of 2017 as similar as what I did for 2016. Then, the findings are combined together to build the prediction model.

This communication is to update the new findings.

Finally, two deliverables were committed as schedule for Supervisor’s overview, including a Rmarkdown data analysis report with findings on the data and a multiple linear regression prediction model to predict 3millions of log error for the properties in California.

