Executive Summary:

The automotive industry is a multi-trillion dollar market with products that serve as a necessity and a luxury to consumers around the world. As the companies that make up the automotive industry evolves, the industry experiences shifts in trends that give consumers a glimpse of what direction the industry is going towards in the future. The purpose of our research is to analyze renowned companies in the industry and discover what technological trends currently exist in the market. We chose six companies and analyzed their annual and 10-k reports from 2013 to 2018. In each annual report, we crawled the most relevant sections to our research goals, cleaned the data to prepare it for topic modeling and analyzed the results. To further increase the value of our analysis, we also looked at how we can determine how innovative a car company is using the results from our topic model. We did this by creating a variable call the innovation score and compared that score to the company's research and development spending. After conducting a thorough analysis, we discovered that there are common trends within the automotive industry. Examples of these trends are solar, electric and energy saving technology. We also found that there is a positive correlation between a company's R&D expenses and how innovative the company is.

Problem Statement:

The goals of this analysis are to discover what technological trends exist in the automotive industry and if there is a correlation between a car company's R&D spending and how innovative the car company is. Car companies in the industry can benefit from this analysis in many ways. Innovation and technology are the driving forces of growth in the industry. In order to remain competitive and grow their business, companies must keep up with the evolving trends. Our analysis provides car companies with information on which specific trends they should invest resources in and if spending more financial resources in R&D increases innovation in their products.

For our hypothesis, we believe that there are common trends in the industry that companies are striving to keep up with. We also think that there is a positive correlation between a company's R&D spending and the company's innovative score. There is also a positive correlation between a company's R&D spending and the company's revenue in the previous year.

Data Description:

To achieve a diverse representation of the automotive industry, we chose six companies that vary in degree of innovation to analyze. These six companies are BMW, Ford, General Motors, Tesla, Toyota, and Volkswagen. For each of these companies, we analyzed content from either their 10-K report or annual report from 2013 - 2018. These reports provide information on many categories pertaining to the company such as management, workforce, R&D, goals and many more. After going through the reports we decided to crawl all reports but only extract certain sections of the reports to the text as they are the most relevant to our goals. The sections that we crawled include the company's business overview, risks, goals and strategy. We also used each company's financial statements to extract information on revenue and R&D expenses for each of those six years.

Methodology:

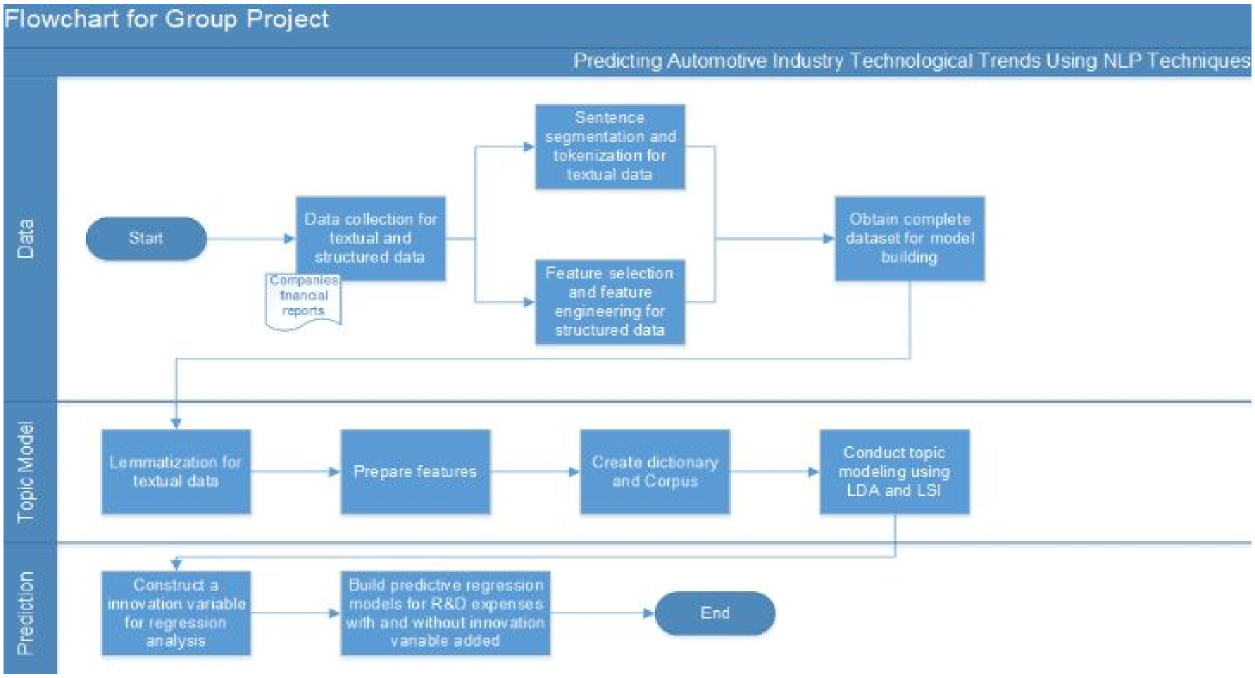

Flowchart for Group Project
Predicting Automotive Industry Technological Trends Using NLP Techniques
Data
Start
Data collection for textual and structured data
Companies financial reports
Sentence segmentation and tokenization for textual data
Feature selection and feature engineering for structured data
Obtain complete dataset for model building
Topic Model
Lemmatization for textual data
Prepare features
Create dictionary and Corpus
Conduct topic modeling using LDA and LSI
Prediction
Construct a innovation variable for regression analysis
Build predictive regression models for R&D expenses with and without innovation variable added
End

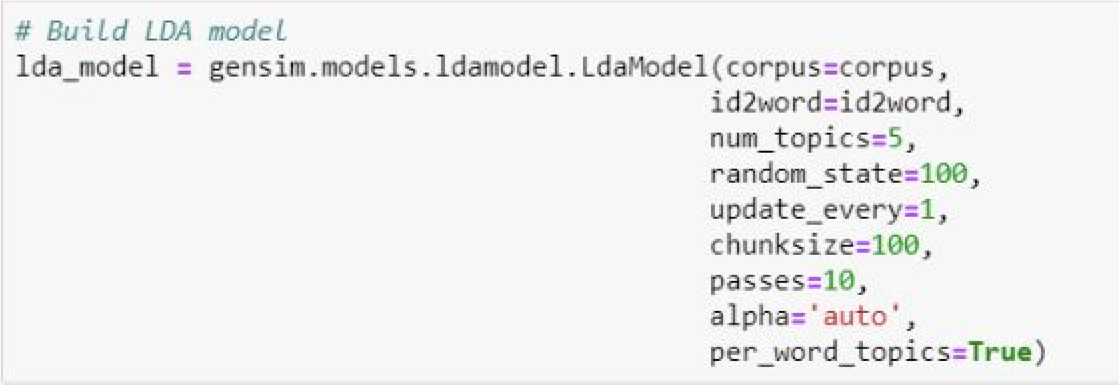
Figure 1: Project Methodology

Figure 1 shows the methodology process of our analysis. After crawling the content from the reports to extract relevant data, we prepared the data for modeling by cleaning the data. To clean the data we first removed the stopwords by importing the *stopwords* list module from *nltk.corpus*. Then we imported the *simple_preprocess* module from *gensim.utils*. This removed all the punctuation and unnecessary characters from the data and tokenized the sentences into a list of words.

The next step was to decide on a model that would produce the best results. We decided on topic modeling as it is a technique that extracts topics from a large amount of text. Specifically, we used Latent Dirichlet Allocation (LDA) to consider the topics distribution within the documents and keywords distribution within the topics to obtain a good composition of topic-keywords distribution, which is a popular algorithm for topic modeling with excellent implementations in Python Gensim package.

Next we tested out both bigram and trigram features to see which would result in better model performance. To measure this, we looked at the coherence and perplexity score. The coherence score measures the quality of the topics generated and the perplexity score measures how well the model is able to predict a sample. Compared to bi-grams, tri-grams had a higher coherence score, and a smaller perplexity score and that is what we went with. The final step in preprocessing our data before running LDA was to lemmatize the text. We imported *spacy*, and ran the code to keep only the nouns, adjectives, verbs and adverbs in our text.

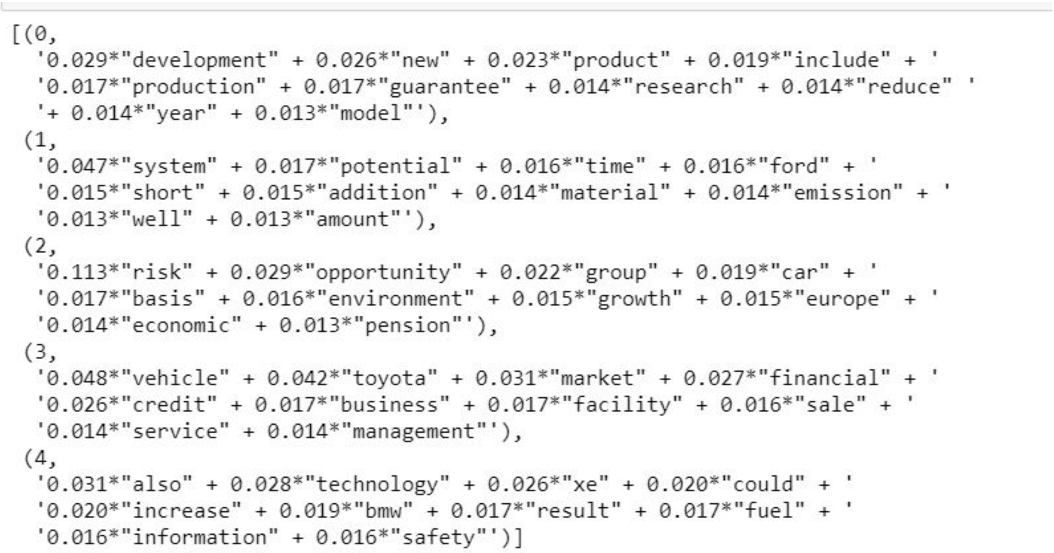
After preprocessing the data, we created the two main inputs of the model which are the dictionary (id2word) and the corpus. The dictionary is created using the data after the text is lemmatized and the corpus is based on the Term Document Frequency. Now that we had the dictionary and corpus, we started building the LDA. For the parameters, we chose to extract five topics, and one hundred as the chunk size. The chunk size is the number of documents to grouped together as a training chunk for the model. Figure 2 shows all the parameters set for the model. We ran the model for each year we collected reports on.



```
# Build LDA model
lda_model = gensim.models.ldamodel.LdaModel(corpus=corpus,
                                           id2word=id2word,
                                           num_topics=5,
                                           random_state=100,
                                           update_every=1,
                                           chunksize=100,
                                           passes=10,
                                           alpha='auto',
                                           per_word_topics=True)
```

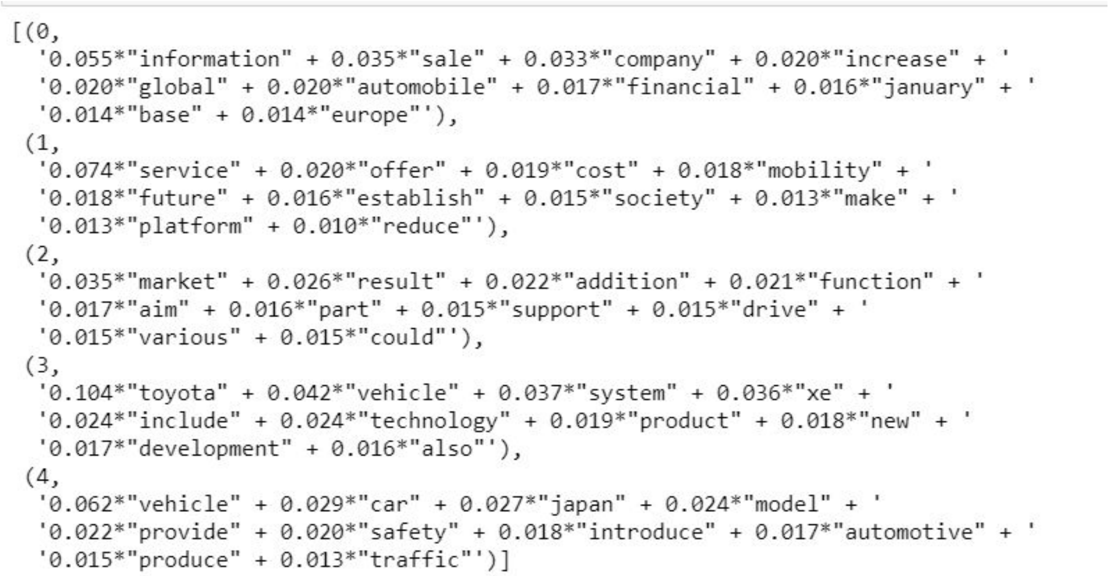
Figure 2: LDA Model Parameters

LDA Model Results:



```
[(0,
  '0.029*"development" + 0.026*"new" + 0.023*"product" + 0.019*"include" + '
  '0.017*"production" + 0.017*"guarantee" + 0.014*"research" + 0.014*"reduce" '
  '+ 0.014*"year" + 0.013*"model"'),
 (1,
  '0.047*"system" + 0.017*"potential" + 0.016*"time" + 0.016*"ford" + '
  '0.015*"short" + 0.015*"addition" + 0.014*"material" + 0.014*"emission" + '
  '0.013*"well" + 0.013*"amount"'),
 (2,
  '0.113*"risk" + 0.029*"opportunity" + 0.022*"group" + 0.019*"car" + '
  '0.017*"basis" + 0.016*"environment" + 0.015*"growth" + 0.015*"europe" + '
  '0.014*"economic" + 0.013*"pension"'),
 (3,
  '0.048*"vehicle" + 0.042*"toyota" + 0.031*"market" + 0.027*"financial" + '
  '0.026*"credit" + 0.017*"business" + 0.017*"facility" + 0.016*"sale" + '
  '0.014*"service" + 0.014*"management"'),
 (4,
  '0.031*"also" + 0.028*"technology" + 0.026*"xe" + 0.020*"could" + '
  '0.020*"increase" + 0.019*"bmw" + 0.017*"result" + 0.017*"fuel" + '
  '0.016*"information" + 0.016*"safety"')]
```

Figure 3: Topics Generated for 2013 Data



```
[(0,
  '0.055*"information" + 0.035*"sale" + 0.033*"company" + 0.020*"increase" + '
  '0.020*"global" + 0.020*"automobile" + 0.017*"financial" + 0.016*"january" + '
  '0.014*"base" + 0.014*"europe"'),
 (1,
  '0.074*"service" + 0.020*"offer" + 0.019*"cost" + 0.018*"mobility" + '
  '0.018*"future" + 0.016*"establish" + 0.015*"society" + 0.013*"make" + '
  '0.013*"platform" + 0.010*"reduce"'),
 (2,
  '0.035*"market" + 0.026*"result" + 0.022*"addition" + 0.021*"function" + '
  '0.017*"aim" + 0.016*"part" + 0.015*"support" + 0.015*"drive" + '
  '0.015*"various" + 0.015*"could"'),
 (3,
  '0.104*"toyota" + 0.042*"vehicle" + 0.037*"system" + 0.036*"xe" + '
  '0.024*"include" + 0.024*"technology" + 0.019*"product" + 0.018*"new" + '
  '0.017*"development" + 0.016*"also"'),
 (4,
  '0.062*"vehicle" + 0.029*"car" + 0.027*"japan" + 0.024*"model" + '
  '0.022*"provide" + 0.020*"safety" + 0.018*"introduce" + 0.017*"automotive" + '
  '0.015*"produce" + 0.013*"traffic"')]
```

Figure 4: Topics Generated for 2017 Data

Topics generated for each company in each year is not listed here due to the space limitation. However, from the results, innovation words like “electricity”, “solar” and “emission” appear with a relatively higher distribution in those topic models relating to each company in each year. This indicates that certain company talks more about those innovative technologies or environmental friendly techniques. The results of two years are displayed in Figures 3 and 4. We can see from the results, that there are mentions related to innovation such as technology and fuel. After generating all the topics for each year, we proceeded to create the factor that would help us test out our hypotheses, the innovation factor for each company.

Innovative Score:

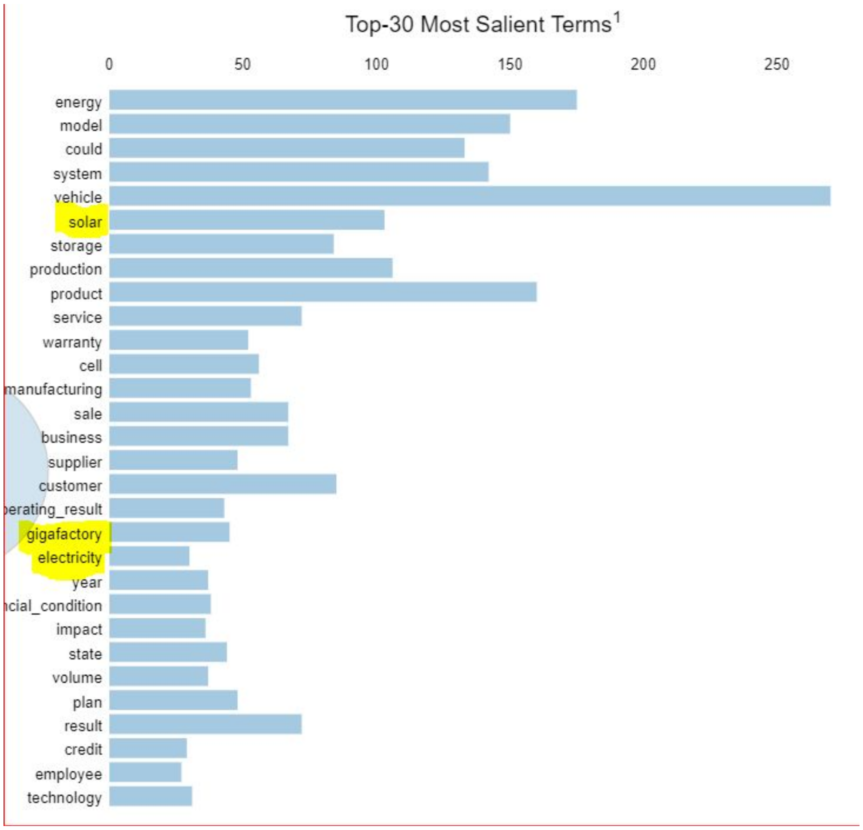


Figure 5: Most Salient Terms with Words Related to Innovation

For each company, we re-ran the model with the same parameters to generate the list of topics found in each company's annual reports. Figure 5 shows an example of terms in the topics generated for one company that are relevant to technology and innovation. To create the innovation score, we looked at the topics in each model to see which had terms associated with innovation. From there we calculated the proportion of topics that were technology associated versus those that were not technology associated and created the weighted variable. From Figure 6, we can see that Tesla has the highest Innovation Score while Ford has the lowest. This means that in all of Tesla's reports, the company mentions words linked to innovation more frequently than the other five companies.

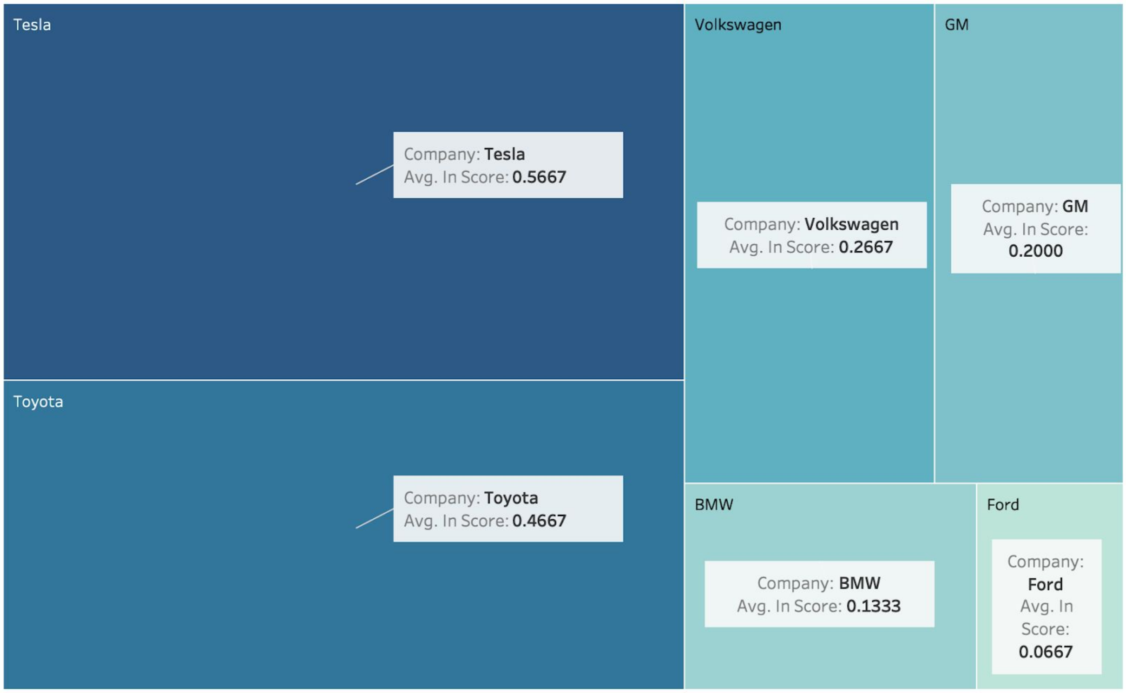


Figure 6: Innovation Score for Each Company

R&D Spending Rate

R&D spending is a crucial component of company's financial planning and budgeting process. We evaluated the in-scope companies' R&D spending over the selected years by dividing the current year R&D spending by its revenue from the previous year. This is consistent with the fact that companies plan for future business goals based on previous year performance.

Figure 7 summarizes quantitatively how different companies invest on R&D initiatives. Tesla has the largest portion of investment on R&D throughout the years. We also noted the R&D spending rate decreased from 2013 (0.559) to 2018 (0.124), which can be partially explained by the fact that a significant amount of setup cost was required in the early stage of the company. We also observed that the R&D rate for other companies tend to stay steady over the years where Toyota invests the least on R&D on average.

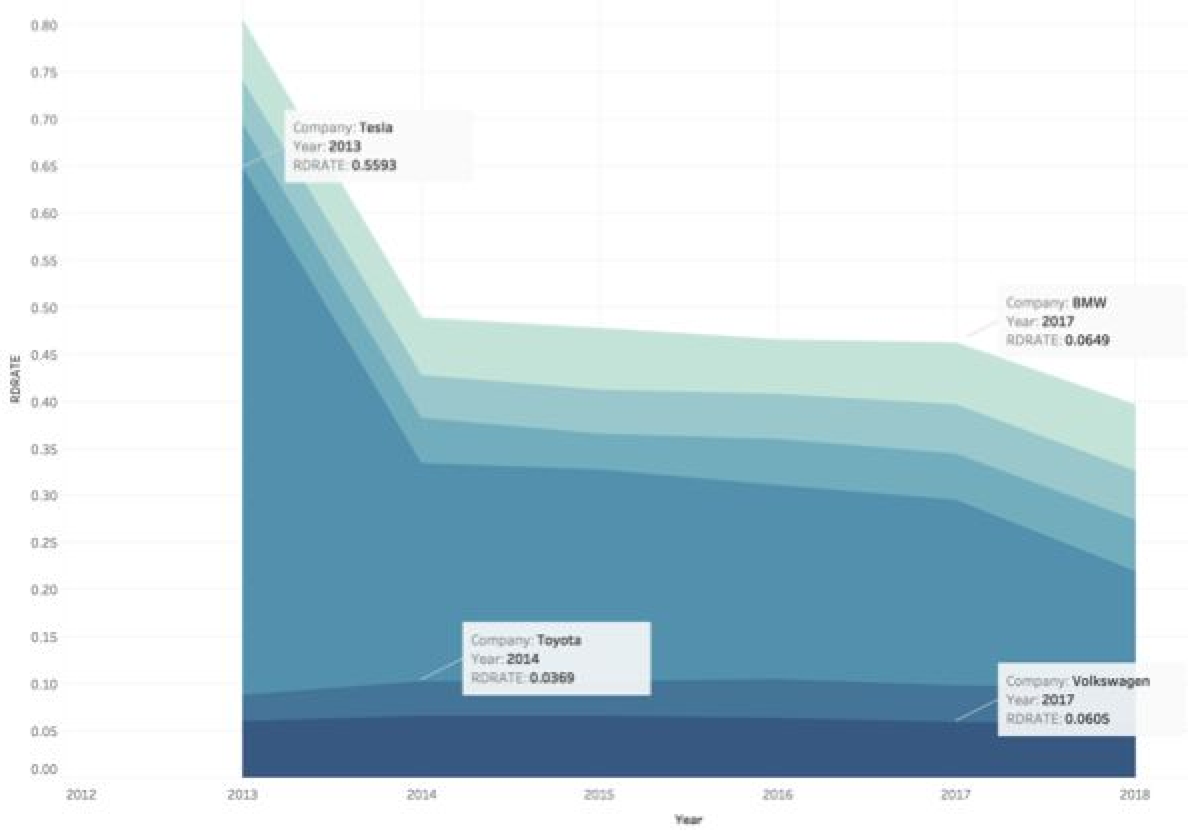


Figure 7: R&D Spending Rate from 2013 - 2018 Summarized by Companies

Topic Modeling Results vs. Financial Data Analysis

Combining the insights drawn from the topic modeling and the financial data analysis, we are able to examine whether the company is investing on innovation as they claim to be. Figure 8 shows that Tesla has both the highest average innovation score also the highest R&D spending rate over the years. Toyota has the second highest innovation score based on its own disclosure in the financial reports. However, its actual spending on R&D is the least among all the companies. This misalignment can be alarming to the stakeholders of the company as it may indicate the financial reporting is not fairly reflecting what the company's current business objectives are.

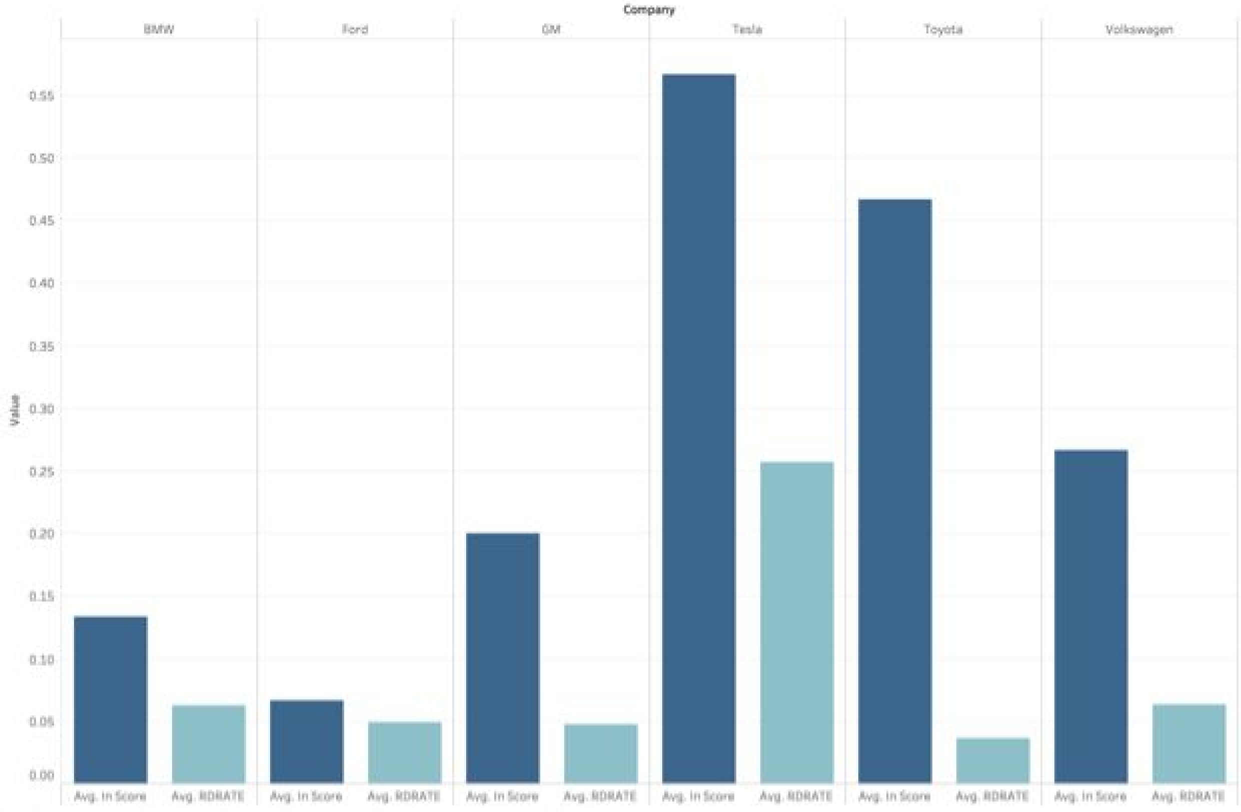


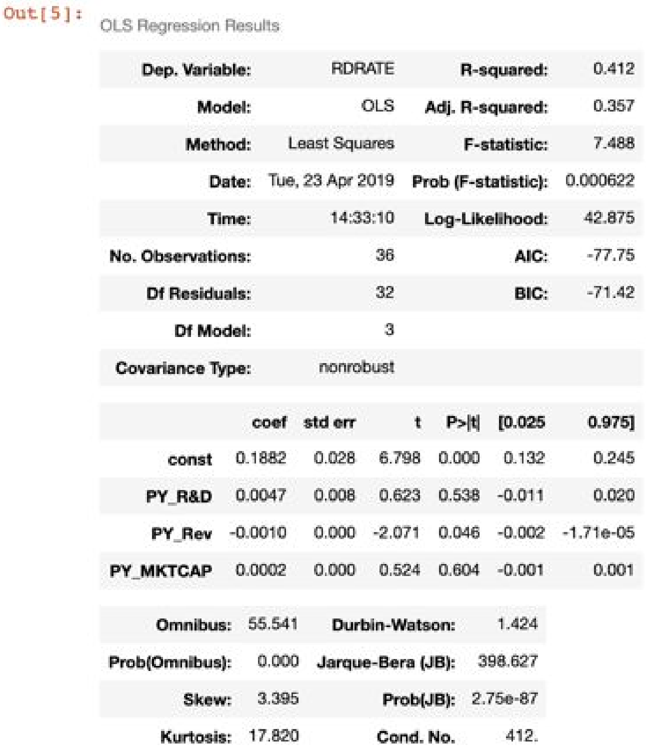
Figure 8: Comparison between Average Innovative Score and Average R&D Rate (CY R&D/PY Revenue) Summarized by Companies

Predictive Model Results

To test our hypotheses, we built two linear regression models and compared their performances in predicting the R&D spending rate for automotive companies. We firstly built a baseline model using previous year's revenue, R&D expense and market capitalization as independent variables, which is in line with the business practice where companies conduct financial planning and budgeting based on previous year's business performance.

The baseline model results in Figure 7 indicates that the adjusted R-squared is 0.357, meaning that 35.7% of the variations in the dependent variable (R&D spending rate) can be explained by the variations in the independent variables.

We subsequently added the innovation scores for each company in each year into the predictive model and obtained the results in Figure 8. It can be observed that the adjusted R-squared increased by 24% to 0.442 after adding the innovation score derived from the topic modeling. Better still, the coefficient of the innovation score variable (0.146) and the associated p-value (0.022) suggests that there is a positive relationship between the company's R&D spending rate and the corresponding innovation score. Thus, we are able to conclude that adding the innovation score significantly improves the model performance to predict the R&D spending rate.



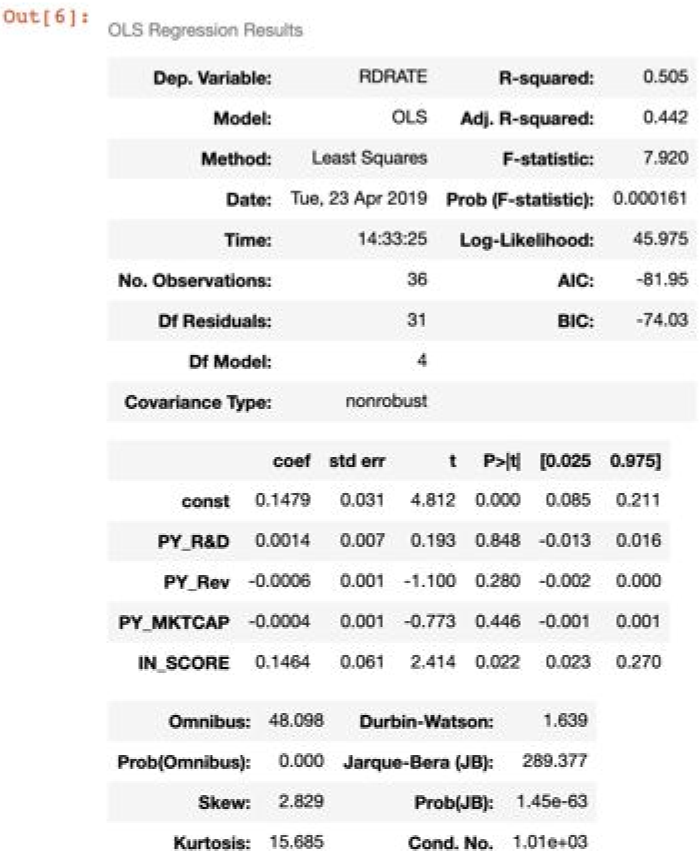
```
Out[5]:
OLS Regression Results

    Dep. Variable:            RDRATE        R-squared:     0.412
           Model:                OLS   Adj. R-squared:     0.357
          Method:      Least Squares      F-statistic:     7.488
            Date:   Tue, 23 Apr 2019 Prob (F-statistic): 0.000622
            Time:           14:33:10   Log-Likelihood:    42.875
No. Observations:                 36              AIC:    -77.75
    Df Residuals:                 32              BIC:    -71.42
        Df Model:                  3
 Covariance Type:          nonrobust

               coef  std err       t   P>|t|  [0.025      0.975]
     const   0.1882    0.028   6.798   0.000   0.132       0.245
    PY_R&D   0.0047    0.008   0.623   0.538  -0.011       0.020
    PY_Rev  -0.0010    0.000  -2.071   0.046  -0.002   -1.71e-05
 PY_MKTCAP   0.0002    0.000   0.524   0.604  -0.001       0.001

      Omnibus:  55.541    Durbin-Watson:     1.424
Prob(Omnibus):   0.000 Jarque-Bera (JB):   398.627
         Skew:   3.395         Prob(JB):  2.75e-87
     Kurtosis:  17.820         Cond. No.      412.
```

Figure 7: Baseline Linear Regression Model Results


Out[6]:

OLS Regression Results

Dep. Variable:	RDRATE	**R-squared:**	0.505
Model:	OLS	**Adj. R-squared:**	0.442
Method:	Least Squares	**F-statistic:**	7.920
Date:	Tue, 23 Apr 2019	**Prob (F-statistic):**	0.000161
Time:	14:33:25	**Log-Likelihood:**	45.975
No. Observations:	36	**AIC:**	-81.95
Df Residuals:	31	**BIC:**	-74.03
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.1479	0.031	4.812	0.000	0.085	0.211
PY_R&D	0.0014	0.007	0.193	0.848	-0.013	0.016
PY_Rev	-0.0006	0.001	-1.100	0.280	-0.002	0.000
PY_MKTCAP	-0.0004	0.001	-0.773	0.446	-0.001	0.001
IN_SCORE	0.1464	0.061	2.414	0.022	0.023	0.270

Omnibus:	48.098	**Durbin-Watson:**	1.639
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	289.377
Skew:	2.829	**Prob(JB):**	1.45e-63
Kurtosis:	15.685	**Cond. No.**	1.01e+03

Figure 8: Linear Regression Model Results with Innovative Score Variable Added

Conclusions

Our research aims to extract insights from both the topic modeling and the predictive modeling. From the topic modeling results, we can firstly identify several technological trends in the automotive industry such as electric vehicles (EVs), autonomous driving and environmental friendly technologies. Subsequently, we ranked the in-scope companies in terms of their innovation score, Tesla was ranked top as a company that endeavors to develop cutting-edge driving technologies. At the same time, Ford was ranked the last, indicating that the company still mainly focuses on traditional automotive related operations.

According to the financial data analysis, we found that Tesla invests the largest proportion of its revenue on R&D initiatives, which stays in line with the insights identified from the topic modeling. On the contrary, despite its intensive topics related to innovation in the financial report, Toyota spends the least percentage on R&D. The misalignment suggests that the company may not be focusing that much on innovation as it claims to be.

Last but not the least, our predictive modeling result shows that the innovation score contributes positively in predicting the company's R&D spending rate. This serves as a motivation for further studies on calculating more companies' innovation score based on their financial reports.

Recommendations

Faced with the ever-changing market dynamics, automotive companies should adopt more data-informed business strategies so that they can catch up with the emerging technological trends. Our research lies the foundation for companies to conduct further analysis regarding their own market position and the competitors' business focus so they can react agilely to the technological dynamics. For instance, given that EVs and autonomous driving technologies are two dominant trends in the industry, traditional automotive companies such as Ford may consider developing related technologies.

In addition, despite the business value our research was able to achieve, companies can potentially expand the data sources to social media, company-related news so that they can obtain more diverse insights of the market dynamics and also the competitor's strategy. Such information can be synthetically applied to guide the company's future business directions.

Reference

Market Capitalization History
https://ycharts.com/companies/VLKPF/market_cap

Gensim Tutorial
https://markroxor.github.io/gensim/tutorials/index.html

SEC Filling Reports
https://www.sec.gov/edgar/searchedgar/companysearch.html