

Comprehensive Projects

By Maggie Wu

Agenda

01

社交媒体产品评论文本分析-Python

02

自动化流程及异常分析(获得集团全球数字化大赛特等奖)

03

数据库及应用设计项目-Oracle Apex

04

金融分析-R语言

05

供应链分析- Python线性规划

/01

社交媒体产品评论文本分析-Python



Data Capture & NLP Project

BANA 780 Problem 3

By Maggie Wu

社交媒体产品评论文本分析-Python :

1. Introduction to the Project

This dataset comes from Twitter by using tweepy to capture 2000 lines of Tesla Model Y reviews with text cleaning processing.

Background:

Tesla is an electric vehicle and energy company in USA, with a market value of US \$210billion, producing and selling electric vehicles; Tesla's vision is to accelerate the global transformation to sustainable energy. Tesla's market share in China is also growing. Tesla can often be seen on the road in recent two years. It has also driven the rapid development of the domestic electric vehicle industry and environmental sustainable development.

In this project, I analyzed the twitter reviews on Tesla Model Y to check for its popularity among English language countries, peoples' sentiments on it, interest in it and main topics they are talking about it.

Datasets:

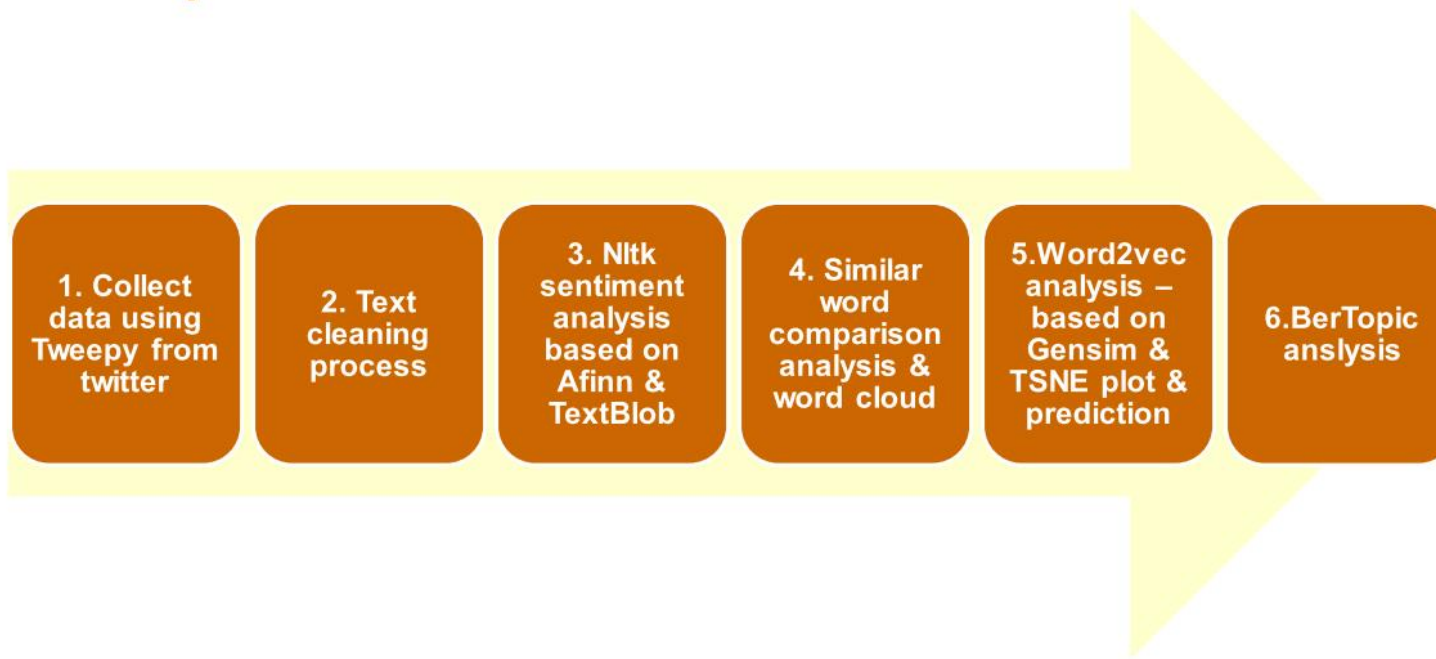
The dataset are mainly about peoples' 2000 recent reviews on Tesla Model Y on in English language countries.

There are 7 dimension as follows:

created_at, text, retweet_count, favorite_count, screenname, location, sentiment, cleantext

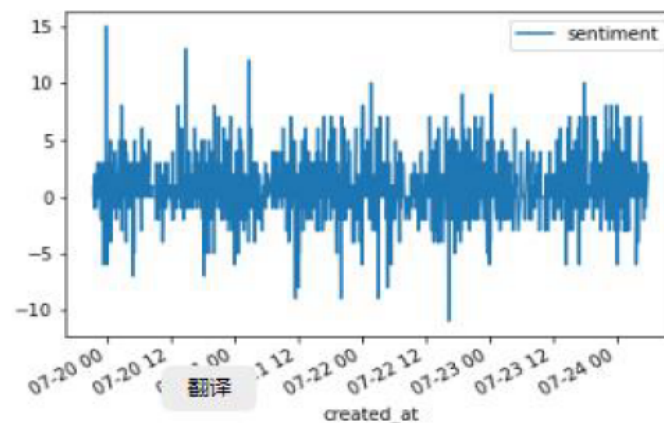
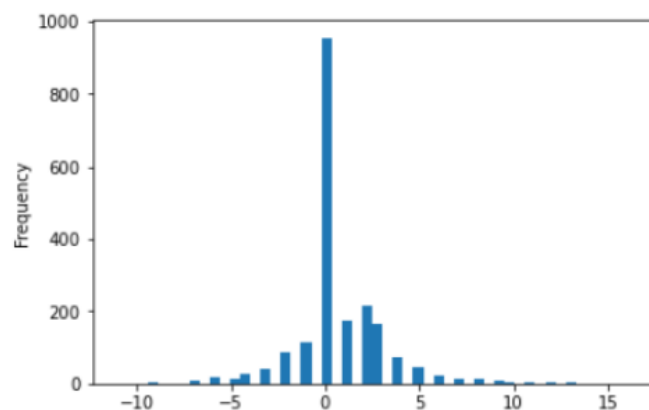
社交媒体产品评论文本分析-Python :

2. Data Manipulation Process



社交媒体产品评论文本分析-Python :

3. Exploring the Unstructured Data – sentiment analysis



Sentiment analysis based on AFINN:

We can find out most sentiment is distributed nearby 0, the positive sentiment is not too high, so as the negative sentiment is not too low. People's sentiment is neutral.

Analyze the sentiment according to time line, the sentiment fluctuation is evenly distributed.

Sentiment analysis based on TextBlob:

Using Tweetblob we can also get the sentiment polarity (-1 ~ +1) is only 0.14 (not high) and objectivity is 0.5.

```
1 tweetblob.sentiment
Sentiment(polarity=0.1416118580308575, subjectivity=0.5031542267732269)
```

3. Exploring the Unstructured Data - common word analysis

```
1 stopwords = nltk.corpus.stopwords.words('english')
2 stopwords.extend(['https', ' ', 'i', 'the', '...', 's', 'it', 'get', 'one', 'my', 'car',
3 wordfreqs = nltk.probability.FreqDist(w.lower() for w in tweetblob.words if w
4 mostcommon = wordfreqs.most_common(30)
5 mostcommon|
```

```
[('tesla', 470),
 ('like', 139),
 ('new', 120),
 ('2022', 78),
 ('would', 74),
 ('battery', 74),
 ('first', 73),
 ('know', 72),
 ('4680', 71),
 ('performance', 66),
 ('giga', 60),
 ('range', 59),
 ('got', 58),
 ('love', 54),
 ('want', 54),
 ('berlin', 52),
 ('think', 52),
 ('really', 50),
 ('made', 49),
 ('production', 47),
```

Using Nltk package to count the most common words

- According to left side, we can see the most common words discussed by people, I think every word indicates one hot topic. We can find people concern about the battery, performance and production. People's sentiment is like, love, want, etc.
- Besides, if someone plan to buy a product and want to be familiar with it, they can use this way to explore:
- For example, the word 4680, when I searched for news about it, I learned it's a kind of battery which has production bottleneck which will lead to production shortage.
- The word giga and berlin, it indicates a Tesla factory Giga Berlin where a Tesla can be manufactured in 45 seconds.

社交媒体产品评论文本分析-Python :

3. Exploring the Unstructured Data - common word analysis

```
: 1 #ngrams = nltk.bigrams(tweetblob.words)
2 ngrams = nltk.bigrams(w.lower() for w in tweetblob.words if w n
3
4
5 ngramfreqs = nltk.probability.FreqDist(ngrams)
6 mostcommon = ngramfreqs.most_common(20)
7 mostcommon

: [(('giga', 'berlin'), 35),
  (('dual', 'motor'), 33),
  (('4680', 'battery'), 28),
  (('tesla', 'dual'), 28),
  (('2022', 'tesla'), 25),
  (('long', 'range'), 18),
  (('battery', 'pack'), 17),
  (('order', 'taiwan'), 17),
  (('tesla', 'giga'), 17),
  (('giga', 'texas'), 17),
  (('chevrolet', 'blazer'), 16),
  (('wan', 'na'), 15),
  (('gon', 'na'), 15),
  (('pay', 'per'), 15),
  (('per', 'minute'), 15),
  (('mustang', 'mach'), 15),
```

Using Nltk package to count the most common 2 related words:

- The left side analysis confirmed my finding on last page. we can find out giga is related to berlin, 4680 is related to battery.
- Other most common 2 related words bring about more new information:
- Per & minute: indicates Tesla's production in high speed;
- Chevolet & blazer: Chevrolet Blazer EV which is the second best seller in USA SUV markets, it's Tesla's strong competitor.

3. Exploring the Unstructured Data – location using word cloud



Word cloud analysis for twitter location:

- In consideration the location is mixed with cities and countries, so I use word cloud to show the distribution.
- Since we use English speaking countries to collect the twitter reviews, so we can see USA, Canada, New York, Australia. However, we can also find out Austria and India people is very interested in Tesla.

社交媒体产品评论文本分析-Python :

3. Exploring the Unstructured Data - similar word analysis

Similar words analysis compared by using different analysis package:

Nltk package:

```
1 text2 = nltk.Text(w.lower() for w in tokens if w not in stopwords)
2 text2.similar('battery')
```

the performance structural from

This analysis is very impressive. Two package show different results:

- When I asked Nltk about battery, they let me pay attention to its performance.
- When I asked Word2vec about battery, they imply 2023 and hope. I think maybe the battery problem can be solved in 2023.

Word2vec package:

```
1 tesla_w2v.wv.most_similar('battery')
```

```
[('2023', 0.3212713599205017),
 ('hope', 0.2639414072036743),
 ('driven', 0.25998568534851074),
 ('used', 0.25749707221984863),
 ('musk', 0.23324371874332428),
 ('ordered', 0.23270562291145325),
 ('yet', 0.22187818586826324),
 ('every', 0.22142019867897034),
 ('tonight', 0.21702739596366882),
 ('buy', 0.20258782804012299)]
```

社交媒体产品评论文本分析-Python :

3. Exploring the Unstructured Data - similar word analysis

Similar words compared by using different analysis package:

Nltk package:

```
1 text = nltk.Text(w.lower() for w in tweetblob.words if w not in stopwords)
2 text.similar('berlin')
```

shanghai perfect differently

- When I asked Nltk about Berlin, they show its competitor Shanghai factory.
- When I asked Word2vec about Berlin, they imply carbon, performance and industrialization. In above mentioned analysis, we have known Giga Berlin's output is 45s/p. Now we got more information that Giga Berlin has adopted more solar energy to reduce carbon emission. Besides, Giga Berlin is trying to reduce the battery cost.

Word2vec package:

```
1 tesla_w2v.wv.most_similar('berlin') ##2
[('carbon', 0.3022310733795166),
 ('make', 0.27051132917404175),
 ('said', 0.26070696115493774),
 ('different', 0.25569725036621094),
 ('coming', 0.22725674510002136),
 ('high', 0.22516673803329468),
 ('making', 0.22131550312042236),
 ('save', 0.21456027030944824),
 ('performance', 0.209019273519516),
 ('industrialization', 0.2072654664516449)]
```

3. Exploring the Unstructured Data - similar word analysis

Similar words compared by using different analysis package:

Nltk package:

```
[72]: 1 text = nltk.Text(w.lower() for w in tweetblob.words if w not in stopwords)
      2 text.similar('problem')
```

- When I asked Nltk about problem, they gave no answer.
- When I asked Word2vec about problem, they show skype, this answer is very interesting. I checked for related news, and found out that Tesla cars will soon support video conferencing (e.g. Skype, and there is a rear view camera in Tesla, which may be used as detecting driver distraction when Autopilot is turned on in future to reduce accidents. Besides, it also mentioned test, which is important for reducing problems.

Word2vec package:

```
1 tesla_w2v.wv.most_similar('problem')
```

```
[('test', 0.9976677298545837),
 ('made', 0.9976261854171753),
 ('car', 0.9975498914718628),
 ('skype', 0.9975183010101318),
 ('first', 0.9975097179412842),
 ('days', 0.997495174407959),
 ('two', 0.9974789619445801),
 ('model', 0.997477650642395),
 ('really', 0.9974758625030518),
 ('would', 0.9974592924118042)]
```

```
: 1 tweetblob.words.count('problem')
```

```
: 12
```

3. Exploring the Unstructured Data –TSNE plot

TSNE Plot Analysis:

- I use TSNE plot to show the relationship between each words (I only show the words which are counted more than 50 times)
- Here I plan to divide them into 4 groups, and made detailed analysis one by one in next pages.



3. Exploring the Unstructured Data –TSNE plot

TSNE Plot Analysis:



Group 1

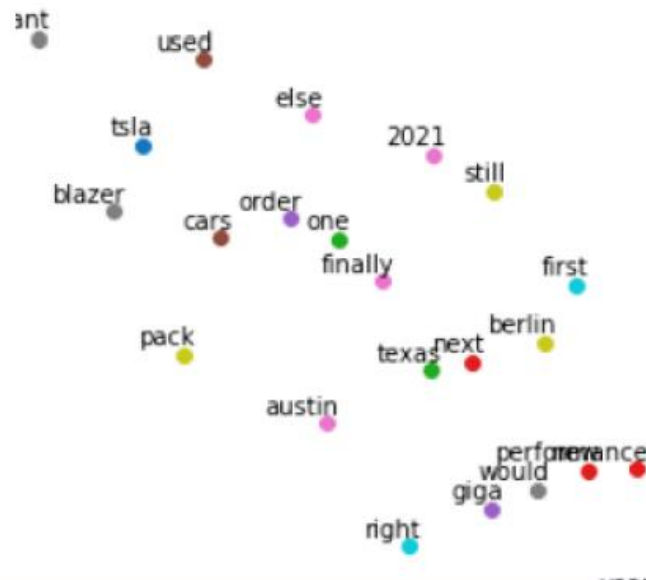
In this group we can see motor, ddpai, battery, 4680, dual, etc. So I think this group is most related to vehicle components and their structure.



社交媒体产品评论文本分析-Python :

3. Exploring the Unstructured Data –TSNE plot

TSNE Plot Analysis:



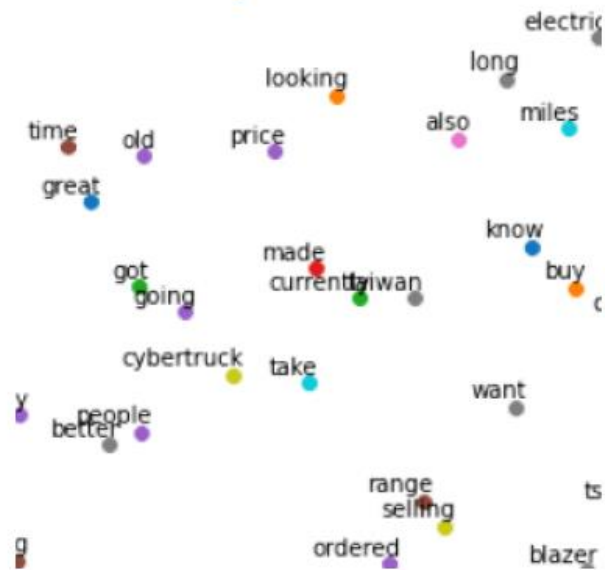
Group 2

In this group we can see blazer (competitor's electric car type), texas berlin, Austin, giga, etc. So I think this group is most related to Tesla's different manufacturing location and its competitors.



3. Exploring the Unstructured Data –TSNE plot

TSNE Plot Analysis:



Group 3

In this group we can see looking, miles, long, price, electric, time, cybertruck, range, selling, etc. So I think this group is most related to people's cost performance requirement towards Tesla's.



Cybertruck

3. Exploring the Unstructured Data –TSNE plot

TSNE Plot Analysis:



Group 4

In this group we can see well, really, love, many, much, production, delivery, months, test, match, etc. So I think this group is most related to Tesla's production and delivery capacity, and people's sentiment (purchasing enthusiasm) and maybe concern about test.



社交媒体产品评论文本分析-Python:

4. Summary and Recommendation

- According to sentiment analysis, out of my expectation, people are not very excited about Tesla, they didn't discuss actively every day. People are calm about it. I think if we can compare with its competitor Chevrolet Blazer EV, the sentiment result would be more clear.
- When I made common words and TSNE analysis, I find it's very helpful for a potential customer, in this way:
 - 1) The potential customer can try to know its competitor as alternative;
 - 2) Pay more attention to its battery problem before purchasing;
 - 3) Giga Berlin is trying to reduce the battery cost, they maybe can wait till 2023 to buy;
 - 4) There may be late delivery problem;
 - 5) They need to take consideration to cost performance.
- According to my analysis, I suggest the potential customer it would be better to wait and buy later.

/02

自动化流程及异常分析

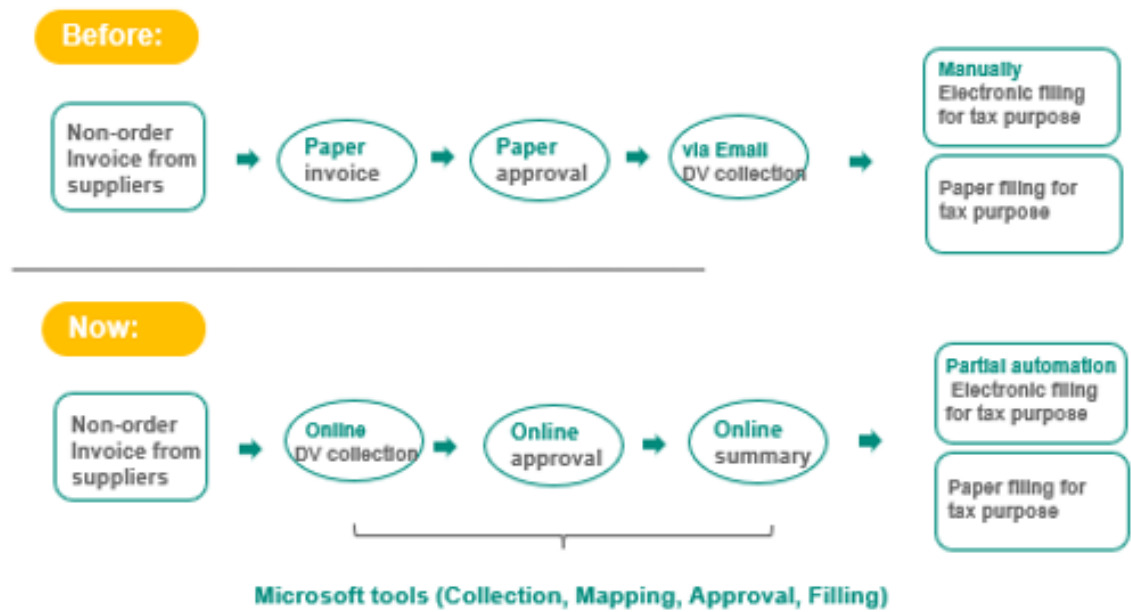


自动化流程及异常分析：

1. Project overview

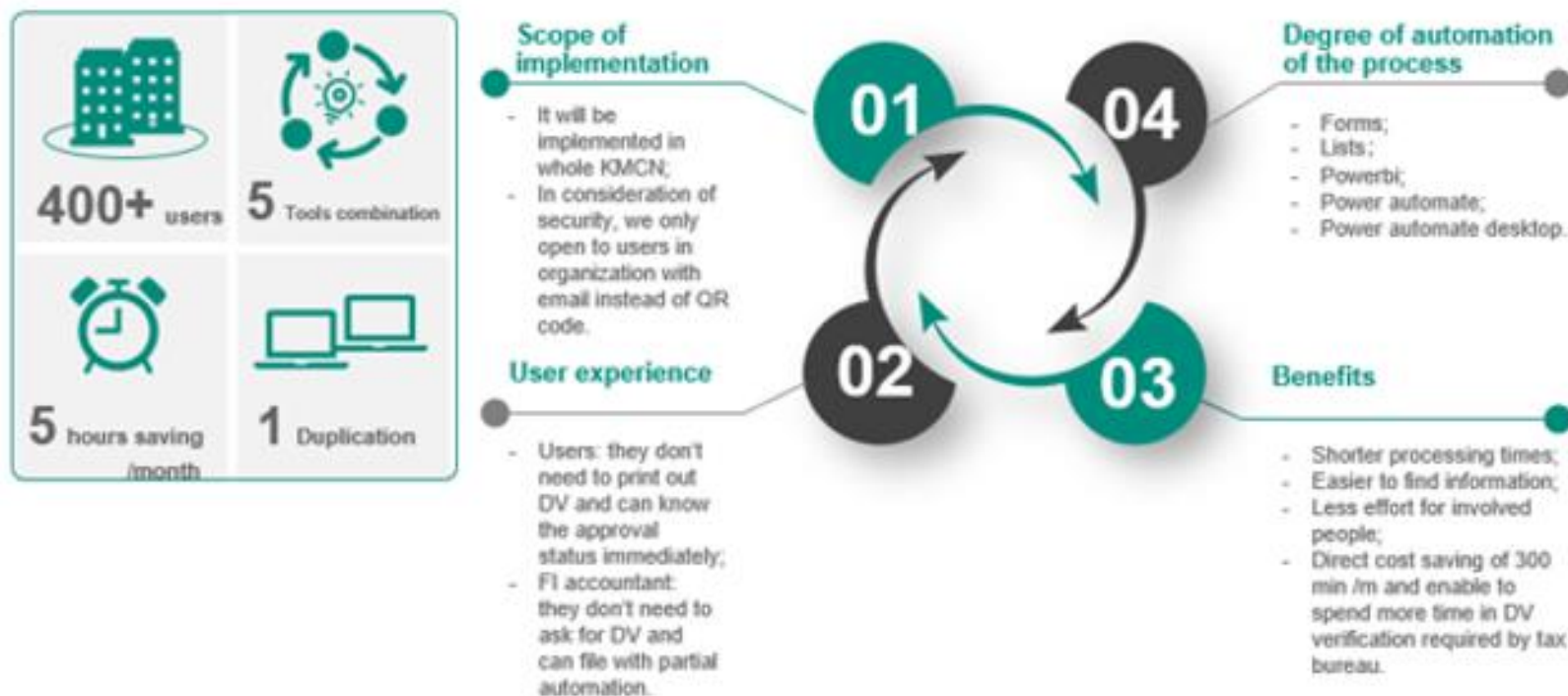
Since China tax bureau began to encourage the usage of digital invoice (DV). Now KMCN receive DV from suppliers who are qualified to issue DV, soon the authority plan to realize the complete DV all over China.

During the current transition period, KMCN need to file DV electronically, and meanwhile file the paper invoice of them, which leads to double workload then before. In order to improve work efficiency, we use Microsoft tools to optimize the workflow as follows.



自动化流程及异常分析:

2. Project details



自动化流程及异常分析:

表单和List工具

CC无订单电子发票报销申请

Compensation application for CC digital invoice without order

Hi, TongXing. When you submit this form, the owner will see your name and email address.

* Required

1. 发票递交人邮箱Invoice submitted Email *

(e.g. nicole.gu@karimayer.com)

Enter your answer

2. 发票含税金额Total invoice value with tax *

(e.g. 250)

Enter your answer

3. 实际报销金额Actual compensation value *

(e.g. 250)

Enter your answer

4. 报销人员工号Employee ID *

Enter your answer

5. 报销人姓名及所在部门Dept. *

Enter your answer

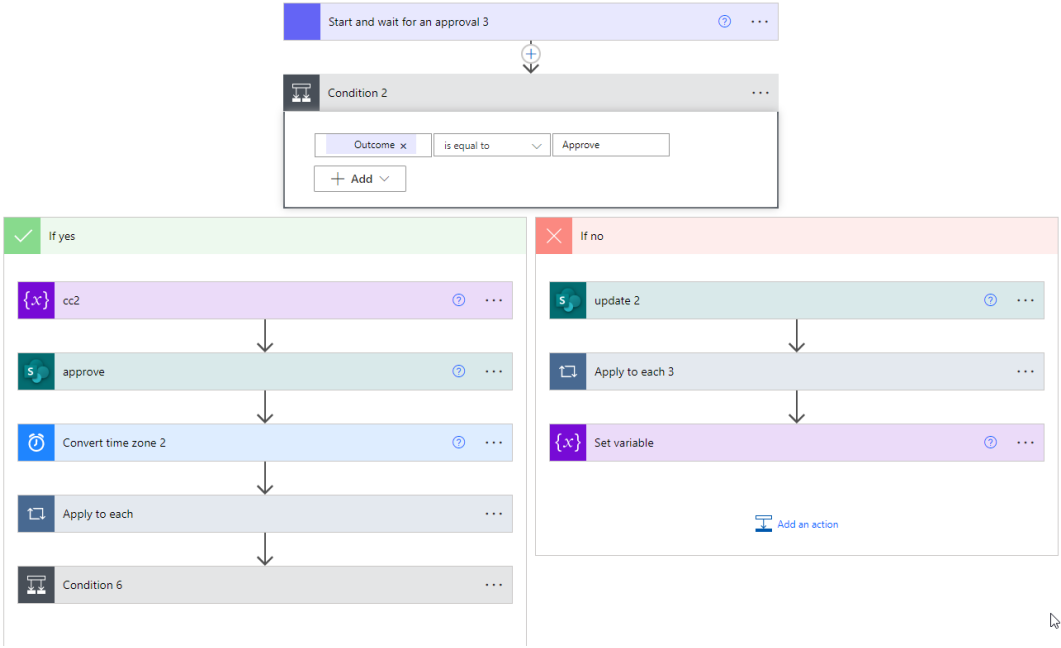
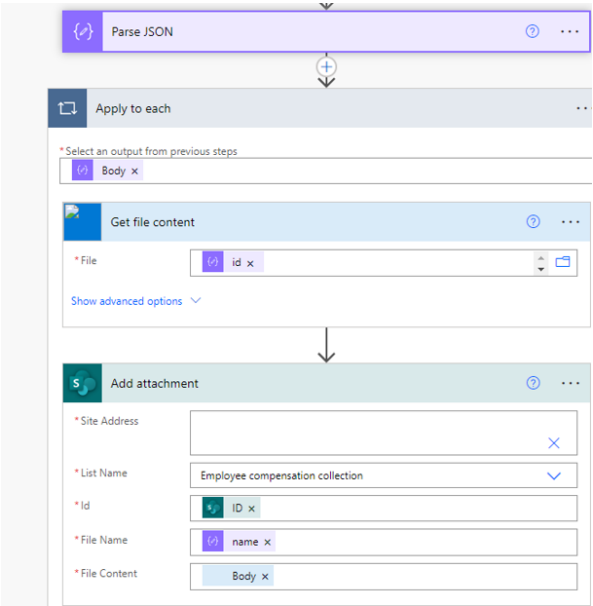
6. 成本中心CC/order号/WBS号 *

Enter your answer

Microsoft Lists										
Search this list										
+ New Edit in grid view Share Export Automate Integrate										
My lists										
Employee compensation collection										
ID	Invoice submitt...	Total invoice va...	Actual compen...	Employee ID	Dept.	CC/order/WBS	Invoice descript...	CC Email	Approval status	sap mark
142		352	352	8010		80-530001	安修		done	Y
143		303	303	080		80-530001	无		rejected	
144		303	303	080		80-530001	无		done	Y
145		275,48	275,48	801		80-726000	部		done	Y
146		250	250	801		80-540001	电		done	Y
147		98	98	801		80-110000	欢		done	Y
148		333,55	333,55	801		80-671002	显		done	Y
149		377,43	377,43	080		80-202000	礼		done	Y
150		880	880	801		80-550001	低		start	
151		141	141	801		80-530001	新		done	
152		201	201	801		80-230000	德		done	
153		79	79	080		80-232002 /681000	电		start	
154		600	600	0801		80-232002/679210	海号		start	

自动化流程及异常分析:

工作流



/03

Oracle Apex

数据库及应用设计项目



数据库及应用设计项目：

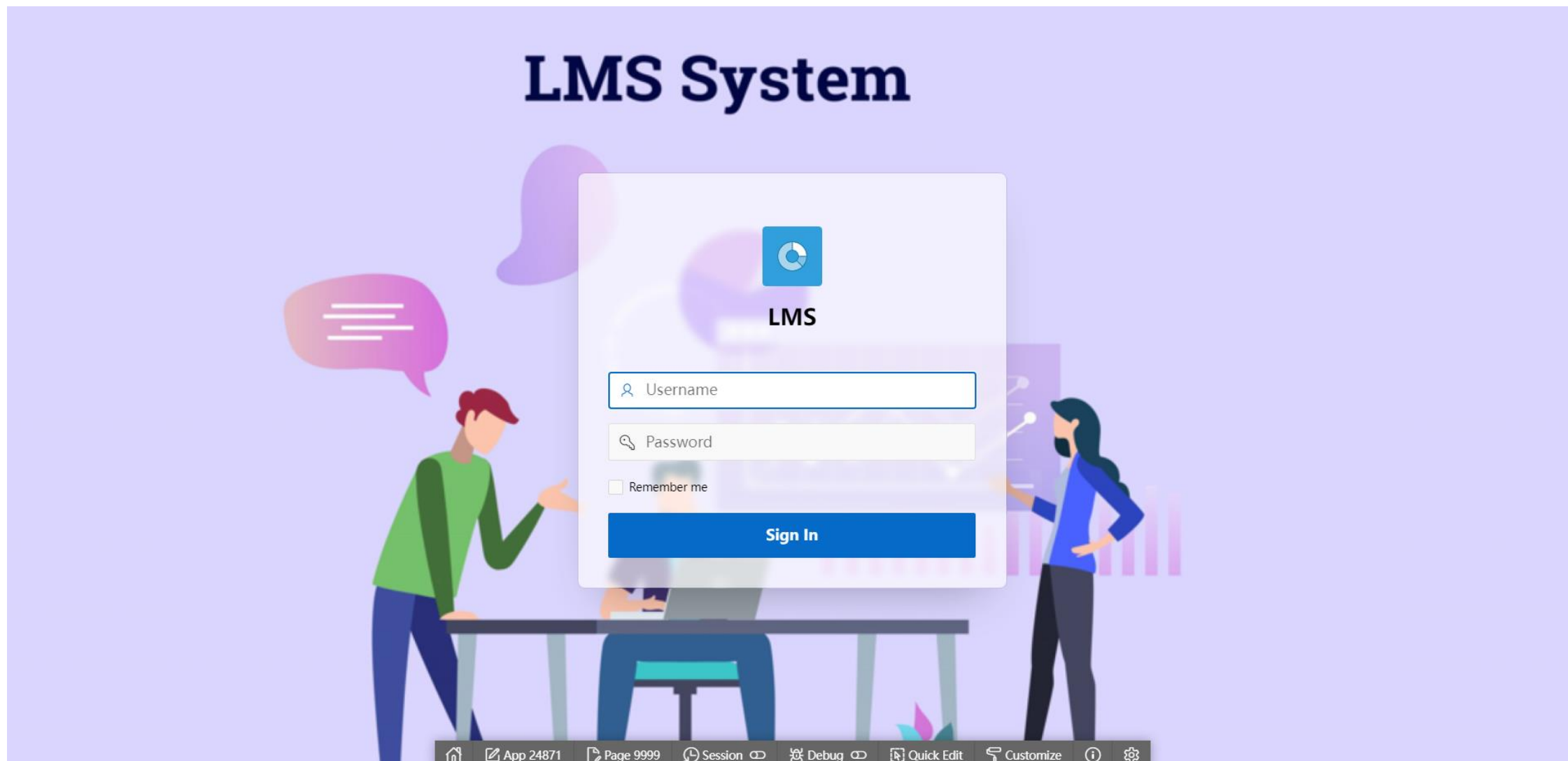
项目介绍：

In today's fast-paced business world, learning and development play a crucial role in the success of any organization. In large organizations, learning management systems (LMS) are used to manage the training and development of employees. However, for many medium and small-sized companies, the cost of such systems can be a barrier to realize this for their employees.

To address this problem, I plan to design a simple learning management application (LMS) that will enable small and medium-sized companies to manage their employee learning progress more effectively and to track the progress to ensure they are meeting company's learning objectives.

To develop this app, I use Oracle SQL to design a database to store the sample data, Oracle PL/SQL to make the trigger for automatically updating and new records ID creation, and Oracle APEX to design the LMS APP by means of the forms, reports, and dashboard function with SQL queries. Besides, I set 2 types of app user role for different authorization. Finally, I made integration test to ensure the app function correctly.

To ensure that the app is user-friendly, I have designed it to be easy to use and navigate. Users will be able to access the necessary forms, reports, and dashboards with just a few clicks. I have also included helpful features, such as search, filter, dashboard functions, to make it easier to find the information needed by users.



数据库及应用设计项目：

×

LMS

Home

Course Setting

Learning Progress

Course Detail

Course Dashboard

Course Selection for Empl...


Achievement Report for E...

Employees Report for HR


Course Report for HR

Achievement Report for HR


Dashboard for HR




LMS




Course Setting



Learning Progress




Course Detail




Course Dashboard


Release 1.0




App 27052




Page 1




Session




Debug




Quick Edit



Customize





↑

数据库及应用设计项目：

LMS

Install App

8012120

Home

Course Setting

Learning Progress

Course Detail

Course Dashboard

Course Selection for Em...

Achievement Report for E...

Employees Report for HR

Course Report for HR

Achievement Report for HR

Dashboard for HR

Course Selection for Employee

Search...

Course Type

Office skills (8)

Personal improvement (3)

Management (2)

Security (1)

Course Length

30 (9)

60 (3)

45 (2)

Supplier

Express (14)

ACC (9)




Cegos (7)

internal (5)

Total Row Count 14

Supplier Express X

Reset

Image	Course Name	Course Length	Course Type	Link	Satification
OneNote Quick StartPoster	OneNote Quick Start	30	Office skills		4.7
	EXCEL data collection	30	Office skills		4.5
	EXCEL Quick Chart	30	Office skills		4.3
Safety production warning filmPoster	Safety production warning film	60	Security		4.3
Getting Started with AccessPoster	Getting Started with Access	30	Office skills		4.3
	Building Your Logical Thinking Using Pyramid Principles	45	Personal improvement		4.2

App 27052

Page 8

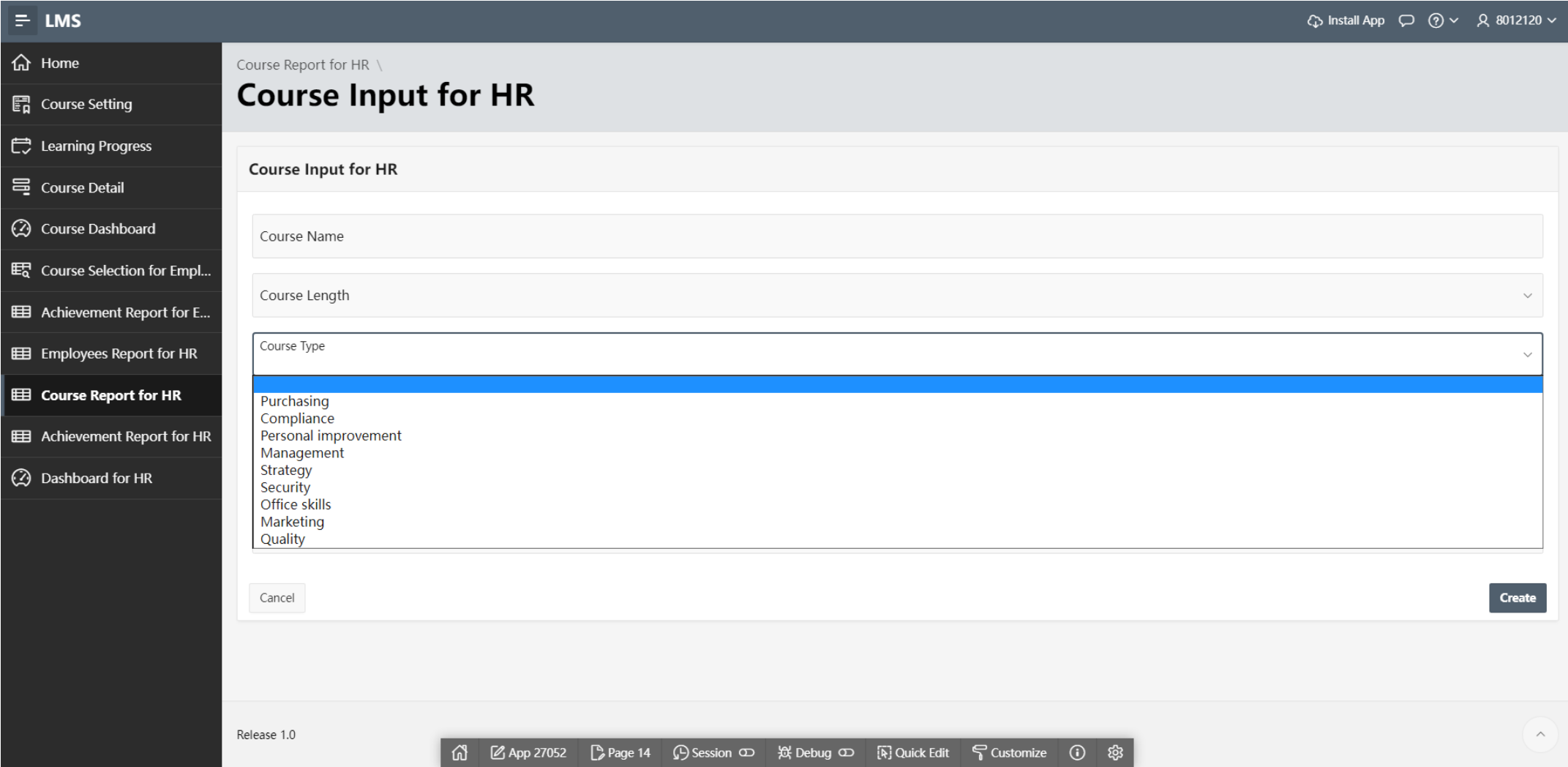
Session

Debug

Quick Edit

Customize

数据库及应用设计项目：



数据库及应用设计项目：

LMS

Home

Course Setting

Learning Progress

Course Detail

Course Dashboard

Course Selection for Empl...

Achievement Report for E...

Employees Report for HR

Course Report for HR

Achievement Report for HR

Dashboard for HR

Employee Report for HR

Employees Input for HR \

Q

Go


Actions

	Employee Name
	Iris
	Johann
	Leah
	Ming
	Allen
	Raven
	Jackie
	Ganlin
	Sally
	Lili
	Cissy
	Kevin
	Pany
	Sandy
	Lenny
	Angel
	Nea

Employees Input for HR

Employee Image

Image2



Employee Name

Leah

Dept Id

Marketing

Image

Choose File

Download

Cancel

Delete

Apply Changes

App 27052Page 17SessionDebugQuick EditCustomize

数据库及应用设计项目：

LMS

Home

Course Setting

Learning Progress

Course Detail

Course Dashboard

Course Selection for Empl...

Achievement Report for E...

Employees Report for HR

Course Report for HR

Achievement Report for ...

Dashboard for HR

Achievement Report for HR

Q

Go

Actions

Create

	Course	Employee	Course Status	Record Date	Satisfaction	Course Selection Type
	Safety production warning film	Iris	Completed	10/11/2022	4	HR administrator
	Safety production warning film	Jackie	Completed	4/19/2022	5	HR administrator
	Safety production warning film	Lili	Completed	10/13/2022	3	HR administrator
	Safety production warning film	Sandy	Completed	4/10/2022	5	HR administrator
	Export Control Compliance Training	Lenney	Completed	2/21/2022	4	HR administrator
	Export Control Compliance Training	Brooke	Completed	3/8/2022	4	HR administrator
	Export Control Compliance Training	Angel	On going	7/11/2022	5	HR administrator
	Compliance training	Neal	Completed	3/2/2022	3	HR administrator
	Compliance training	Cissy	Completed	3/2/2022	5	HR administrator
	Compliance training	Alberto	Completed	3/2/2022	3	HR administrator
	Compliance training	Miya	Completed	3/2/2022	5	HR administrator
	Compliance training	Bella	Completed	3/2/2022	5	HR administrator
	Compliance training	Charles	Completed	3/2/2022	4	HR administrator
	Compliance training	Fiore	On going	4/15/2022	4	HR administrator
	Compliance training	Lina	On going	7/7/2022	4	HR administrator
	Compliance training	Jerry	Completed	4/22/2022	5	HR administrator
	Compliance training	Anna	Completed	4/22/2022	5	HR administrator
	Compliance training				3	HR administrator

App 27052

Page 15

Session

Debug

Quick Edit

Customize

22

数据库及应用设计项目：



/04

金融分析-R语言



Financial Analysis with R 金融分析:

项目介绍: Portfolio Analyzing & Optimizing

As the global economic trend is affected by multiple factors such as politics, global trade, and climate, the financial investment industry is facing enormous challenges. In recent years, quantitative trading based on financial big data and data science has been widely used in the investment field. Quantitative investment is an investment technology that integrates big data science, statistics, economics, finance and information science. Through this project, we master how to use R language to calculate the optimal investment portfolio.

1) By Sharpe Ratio

The Sharpe ratio is a measure of return often used to compare the performance of investment managers by making an adjustment for risk. People can compare investments and assess the amount of risk that each one has per percentage point of return. This helps people better control their risk exposure. The higher the rate, the more returns the investment offers relative to the risks involved. Firstly, calculating mean(profit) and variance(risk), and then using a random sampling of portfolio weights, finally by find the point of tangency (the risk free line tangent to the efficient frontier), which is also with the highest Sharpe Ratio;

2) By Sortino Ratio

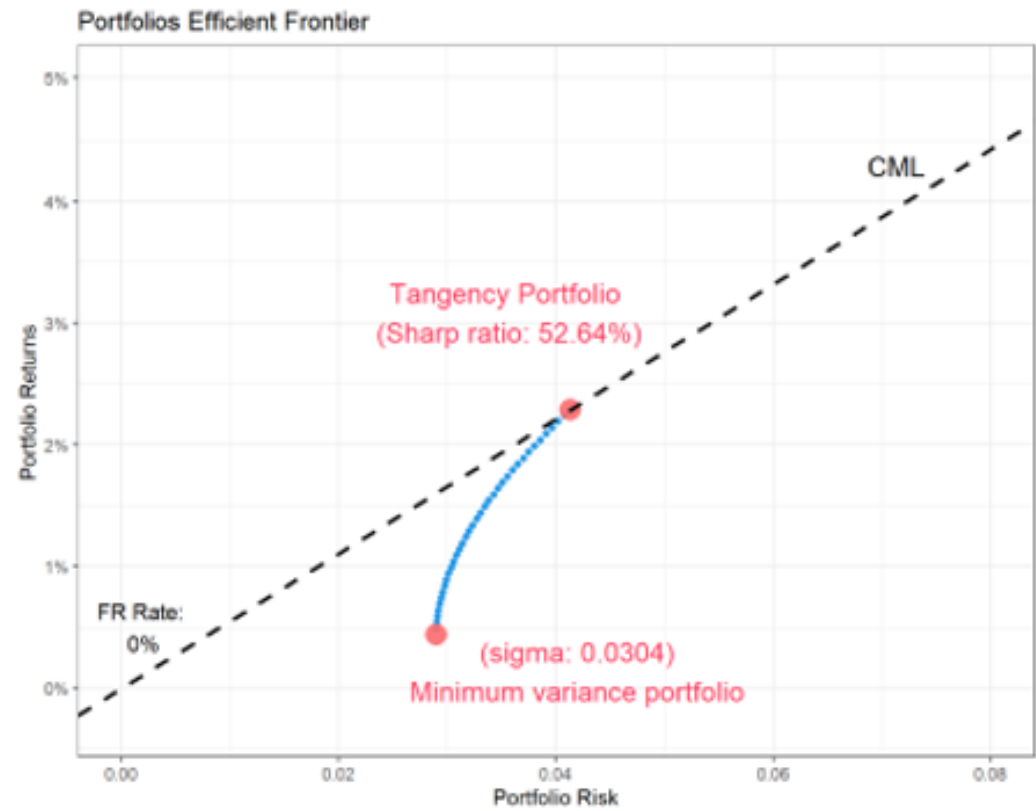
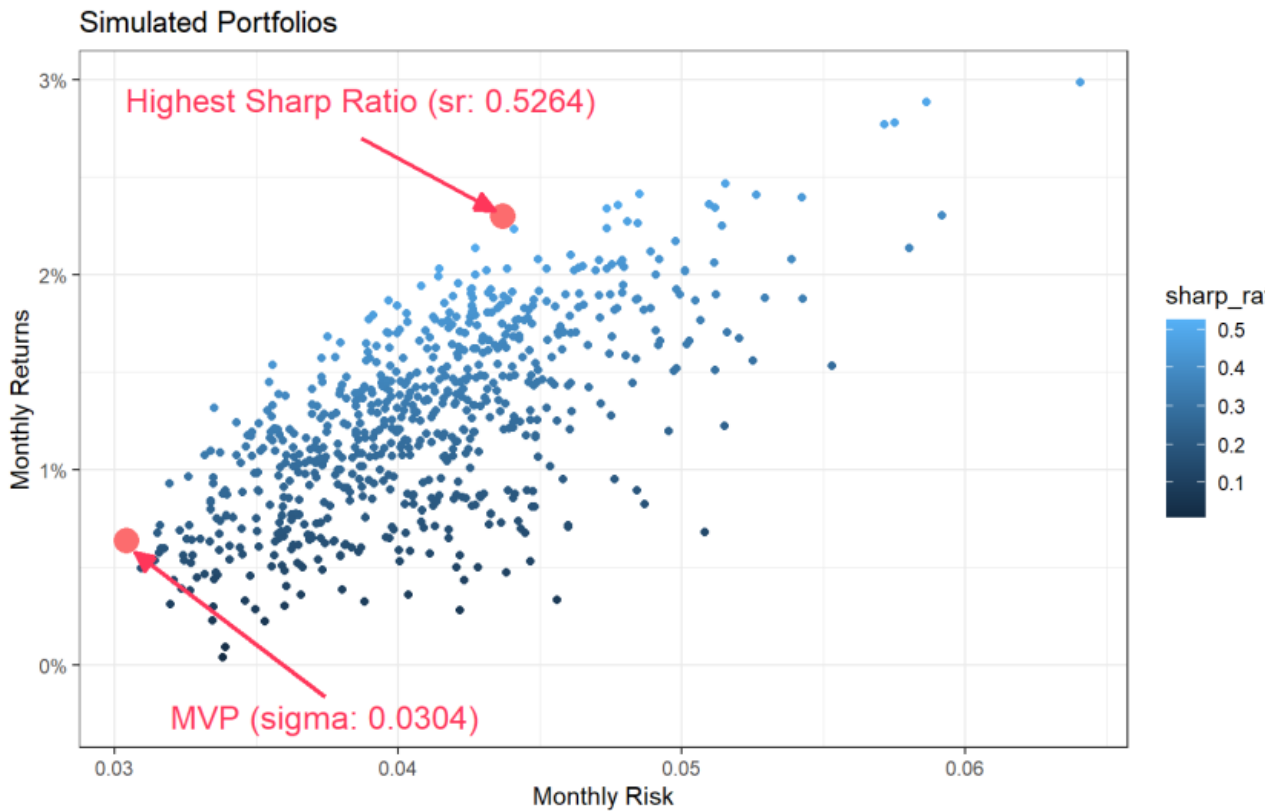
It uses downside deviation instead of standard deviation in Sharpe Ratio, that is measure the risk only returns falling below a specified target, while above target are set to zero.

3) Quadratic Optimization

By using package PortfolioAnalytics we can design the optimal portfolio according our demand, e.g. to let weight add to 1, set the portfolio return to certain target and set the individual weight limit.

Financial Analysis with R 金融分析:

分析结果: Portfolio Analyzing & Optimizing



/05

供应链分析- Python线性规划



供应链分析- Python线性规划:

项目介绍:

- Where to set up the Plant to realize the minimum cost?
- Which plant needs to be closed due to higher cost?

Modeling

- Production at regional facilities
 - Two plant sizes (low / high)
- Exporting production to other regions
- Production facilities open / close



供应链分析- Python线性规划：

Coding Result:

In [76]:

```
## ppt 案例 case study
import pulp
# Initialize Class
model = pulp.LpProblem("Capacitated Plant Location Model", pulp.LpMinimize)
# Define Decision Variables
locc = ['USA','Germany','Japan','Brazil','India']
size = ['Low_Cap','High_Cap']

# Supply_Region
# USA          2719.6
# Germany       84.1
# Japan        1676.8
# Brazil       145.4
# India        156.4

# demand={'USA': 2719.6,
#          'Germany': 84.1,
#          'Japan': 1676.8,
#          'Brazil': 145.4,
#          'India': 156.4}

demand={'USA': 'Dmd': 2719.6,
        ('Germany','Dmd'): 84.1,
        ('Japan','Dmd'): 1676.8,
        ('Brazil','Dmd'): 145.4,
        ('India','Dmd'): 156.4}

fix_cost={'USA': 'Low_Cap': 6500,
          ('Germany','Low_Cap'): 4980,
          ('Japan','Low_Cap'): 6230,
          ('Brazil','Low_Cap'): 3230,
          ('India','Low_Cap'): 2110,
          ('USA','High_Cap'): 9500,
          ('Germany','High_Cap'): 7270,
          ('Japan','High_Cap'): 9100,
          ('Brazil','High_Cap'): 4730,
          ('India','High_Cap'): 3080}

('Japan','India'): 9,
('Brazil','USA'): 12,
('Brazil','Germany'): 14,
('Brazil','Japan'): 21,
('Brazil','Brazil'): 8,
('Brazil','India'): 21,
('India','USA'): 22,
('India','Germany'): 13,
('India','Japan'): 10,
('India','Brazil'): 23,
('India','India'): 8)

cap={'USA': 'Low_Cap': 500,
     ('Germany','Low_Cap'): 500,
     ('Japan','Low_Cap'): 500,
     ('Brazil','Low_Cap'): 500,
     ('India','Low_Cap'): 500,
     ('USA','High_Cap'): 1500,
     ('Germany','High_Cap'): 1500,
     ('Japan','High_Cap'): 1500,
     ('Brazil','High_Cap'): 1500,
     ('India','High_Cap'): 1500}

x = pulp.LpVariable.dicts("production",[(i,j) for i in locc for j in locc],
                           lowBound=0, upBound=None, cat='Continuous') #Integer
y = pulp.LpVariable.dicts("plant",[(i,s) for s in size for i in locc], cat='Binary')
# Define objective function
model += (pulp.lpSum([fix_cost[(i,s)]*y[(i,s)] for s in size for i in locc])
          + pulp.lpSum([var_cost[(i,j)]*x[(i,j)] for i in locc for j in locc]))

# for c in locc:
#     model += pulp.lpSum([cap[(i, s)] for s in size for i in locc ]) == demand[c,

# Define the Constraints
for j in locc:
    model += pulp.lpSum([x[(i, j)] for i in locc]) == demand[j,'Dmd'] #demand[j,
for i in locc:
    model += pulp.lpSum([x[(i, j)] for j in locc]) <= pulp.lpSum([cap[(i, s)]*y[(i,
                           for s in size)])
```

```
# Solve Model
model.solve()
# 6. 打印结果
print(model) # 输出问题设定参数和条件
print("优化状态:", pulp.LpStatus[model.status])
for v in model.variables():
    print(v.name, "=", v.varValue)
print("最优总成本 = ", pulp.value(model.objective))
```

```
production_({'USA','India'}) Continuous
production_({'USA','Japan'}) Continuous
production_({'USA','USA'}) Continuous
```

```
优化状态: Optimal
plant_({'Brazil','High_Cap'}) = 1.0
plant_({'Brazil','Low_Cap'}) = 0.0
plant_({'Germany','High_Cap'}) = 0.0
plant_({'Germany','Low_Cap'}) = 0.0
plant_({'India','High_Cap'}) = 0.0
plant_({'India','Low_Cap'}) = 1.0
plant_({'Japan','High_Cap'}) = 1.0
plant_({'Japan','Low_Cap'}) = 0.0
plant_({'USA','High_Cap'}) = 1.0
plant_({'USA','Low_Cap'}) = 0.0
production_({'Brazil','Brazil'}) = 145.4
production_({'Brazil','Germany'}) = 0.0
production_({'Brazil','India'}) = 0.0
production_({'Brazil','Japan'}) = 0.0
production_({'Brazil','USA'}) = 1219.6
production_({'Germany','Brazil'}) = 0.0
```



Thank you!