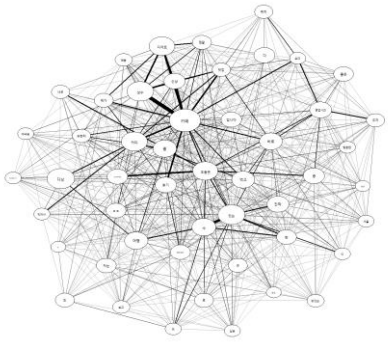
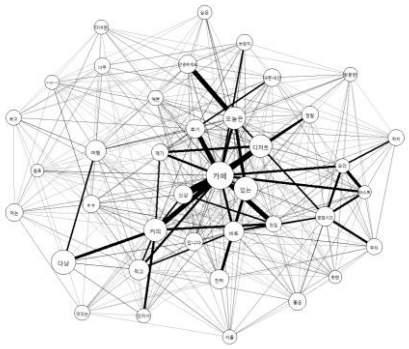
1. 크롤링 후 네트워크 분석 시작

일단 먼저 "카페"에 대해 크롤링을 진행 하여 네트워크 분석으로 사람들이 카공족에 대한 인식이 어떤지 살펴보았습니다.

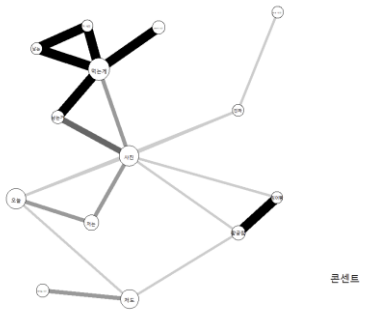


<- 전처리 전



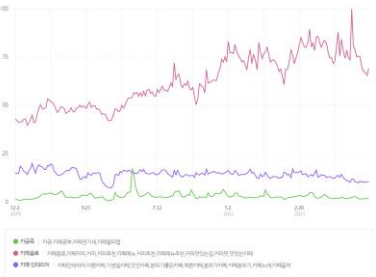
<- 전처리 후

처음 카페 분석을 했을때는 전처리가 되지 않아서 정보를 골라내기 힘들었습니다. 따라서 쓸모없는 데이터를 지우고 필요해보이는 데이터를 덧붙이는 작업을 추가해 나갔습니다. 하지만 여전히 여행, 커피맛, 영업시간, 위치 주차장등 카공족에 대한 키워드는 찾아볼 수 없었고 따라서, 카공족에 대한 사람들의 불편도는 생각보다 크지 않음을 알 수 있었습니다. 또한, 사용자가 아닌 사장님들의 불편도를 알아보기 위해서 카페사장님들이 모여 있는 네이버 카페인 '전국카페협동조합' 중 인증받은 카페사장님들만 들어올 수 있는 자유게시판(인증)코너에서 4개월간 올라온 약 800개의 글을 크롤링을 시작해 보았습니다.



콘센트

15322	흠??? 카공족이 준것 같은 느낌이? [14]	오늘도	2023.03.28.	70
15144	카공족뉴스 [4]	파니아	2023.03.24.	46
15131	카공족 관련한 비디오머그 입니다. [9]	스노뽀두	2023.03.24.	60
14942	즈집은 카공족도 없는데... [10]	토파즈	2023.03.17.	95
14928	카공족 관련해서 인터뷰 해주실 사장님~~ [2]	전국카페사장연합회장	2023.03.16.	106
14510	카공족에 대한 반응 [17]	소확행	2023.02.25.	110
14105	펌)카공족에 대한 생각 [21]	소확행	2023.02.11.	121
5025	당분간은 카공족도 잘 챙겨야..;; [34]	이팀장	2021.07.09.	254
4829	카공족..ㅠㅠ [27]	정직한너의팻	2021.06.22.	172
3014	혼자 오신 카공족... [14]	나를사랑한너	2021.02.13.	214
2070	아...카공족....1인손님.... [3]	Joymom1204	2021.01.18.	460

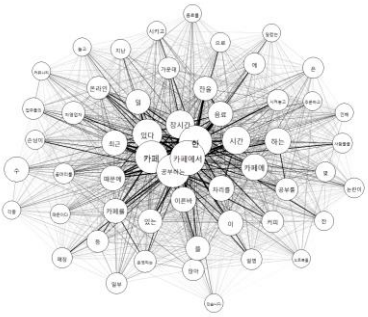


- 카공족 이미지: 중립적(45.4%) 혹은 부정적(39.5%)
- 10명 중 2-3명, 공부 등의 목적으로 카페 방문 경험 있어(29.4%)
--> 공부하면서 지킨 에티켓은
(1) 테이블 최소한으로 차지하기(86.6%)
(2) 손님이 없는 한가한 시간 이용하기(62.5%)
(3) 핸드폰/노트북 충전기 자체적으로 해결하기(56.8%)
(4) 일정 시간이 지나면 추가 주문하기(52.5%)
- 앞으로 공부하러 카페에 장시간 머무를 의향이 있다: 18%
- 카공족 규제 동의도
(1) 테이블 최소한으로 차지하기(85.7점)
(2) 손님이 없는 한가한 시간 이용하기(84.7점)
(3) 일정 시간이 지나면 추가 주문하기(81.3점)
(4) 핸드폰/노트북 충전기 자체적으로 해결하기(79.8점)

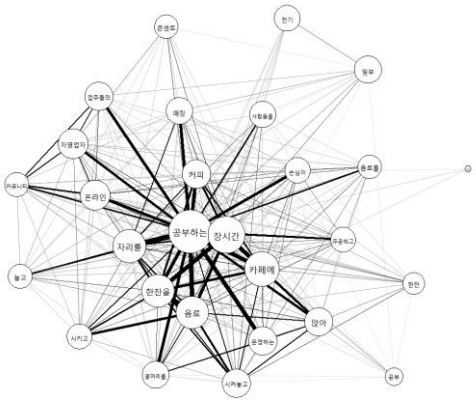
하지만 전기, 카공족, 카페 등과 관련된 단어는 마찬가지로 전혀 찾을 수 없었으며, 유일하게 찾은 콘센트 라는 단어도 오래된 콘센트 점검해 보라는 문장 하나가 전부였습니다. 실제로 카공족 관련 언급은 9개월간 단 1건도 없었으며, 전체 게시글 약 11300개 중 2023년 총 7건, 2021년 4건

이 전부였습니다. 또한, 네이버 트렌드 검색 결과도 카공족은 카페인테리어, 카페 음료 맛에 비해 확연히 검색결과가 떨어지며, 소비자들도 카공족에 따라 카페를 선정하지는 않는 것 같다는 결론을 내릴 수 있었습니다. 종합적으로 사장님들의 카공족에 대한 고민은 크지 않은 것으로 나타났으며 사업성이 충분하다고 느껴지지는 않았습니다. 그래도 '두잇서베이'에서 진행한 연구에 따르면 카공족에 대한 인식도는 부정적이 39.5%로 좋지 않은 것으로 나타났으므로 이 기회를 들어 카공족에 대한 인식을 재고해볼 수 있는 기회로 여겨볼 수 있지 않을까 라는 생각이 들어 조사는 계속해 보기로 했습니다.

먼저, 카공족에 대한 사업에 관한 연구를 계속해 보기 위해 "카공족"이라는 단어를 특정지어서 네이버 뉴스 크롤링을 해보았습니다.



<-전처리전



<-전처리 후

먼저 그냥 네트워크 분석을 해본 결과 유의미한 특징을 잡을 수 없었고, 약간의 전처리 과정을 거쳐 네트워크 분석을 해본 결과 [카페에, 장시간, 음료, 한잔을, 시켜놓고, 사람들을, 골머리를, 자리] 등의 키워드가 연관되어 있는 것을 볼 수 있었습니다. 이를 통해 카페 사장 입장에서 음료 한잔 시키고 자리를 차지해 사람들이 자리에 못앉는 경우에 골머리를 썩고있다고 추론해 볼 수 있었습니다. 또한 [전기, 콘센트]등의 키워드가 여러가지 키워드랑 연관되어 있는것으로 보아 콘센트를 많이 사용하거나 전기를 많이 사용하는 경우 문제가 되고 있음을 알 수 있었습니다.

2. 네트워크 분석 후 가정 도출

따라서 데이터 분석 결과에 따른 가정을 다음과 같이 세웠습니다.

- 카페에서 혼잡한 시간에 오랫동안 자리를 차지하여 다른 사람들이 못 앉게 만드는 사람들이 문제다
- 카페에서 음료 한잔을 시켜놓고(매출에 도움이 되지 않고) 자리를 차지하는 사람들이 문제다.
- 카페에서 전기를 많이 쓰는 손님들 때문에 카페 운영 측면에서 부담이 될 수 있다.

또한 데이터에는 나타나지 않지만 개인적인 가설 또한 다음과 같이 세웠습니다.

- 자주 방문해주는 사람의 경우 사장님과 내적 친밀도가 쌓여있기 때문에 그사람에 대한 사장님의 만족도는 높을 것이다.

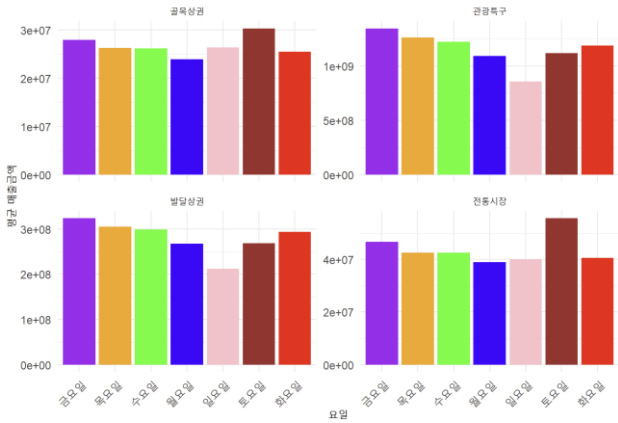
이에 대해 셰빌리티 사장님이 구상중인 '전기사용 시간에 따라 돈을 받겠다' 라는 아이디어는 크게 '카페가 얼마나 혼잡한가' 그리고 '카페에서 전기를 얼마나 많이 썼는가', '카페에서 매출을 얼마나 내주었는가', '얼마큼 단골 손님인가' 에 따라서 돈을 다르게 받아야 카페 사장님들의 만족도가 가장 올라갈 것으로 생각이 됩니다.

3. 카페 혼잡도 데이터 구성

따라서 먼저 카페 혼잡도를 계산해보기 위해 서울 열린데이터 광장에서 제공하는 '서울시 상권분석 서비스 데이터_2022년.csv'를 가져왔습니다. 서울시 상권분석 서비스 데이터에는 요일별, 그리고 시간대별 매출 그리고 각 매출이 나오게 된 상권에 대한 특징이 담겨져 있었습니다. 다음은 요일별 그리고 시간대별 매출 데이터에 대한 결과입니다.

요일별 데이터

	월요일	화요일	수요일	목요일	금요일	토요일	일요일
골목상권	23922374	25536555	26167531	26289187	28005837	30330937	26371575
관광특구	1257490773	1092177273	1185367348	1218401412	1342210505	1117585076	856152560
발달상권	267355686	294066116	299023704	305513360	323814361	268734050	212522026
전통시장	38894790	40641967	42637150	42544452	46780791	55687101	40055414



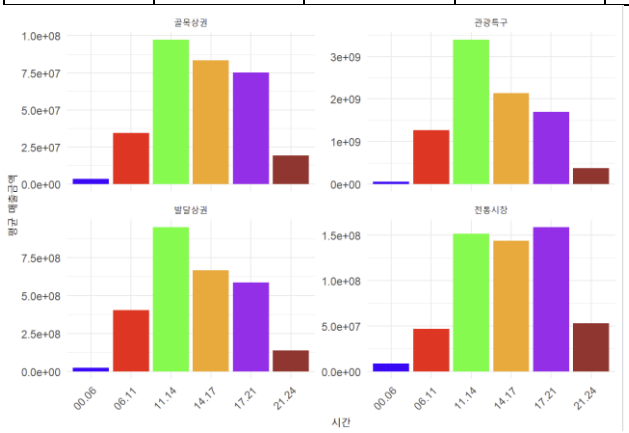
셰빌리티 사장님이 공수해오신 자료의 가게들의 위치를 찍어보니 대부분 발달상권에 위치한 경우가 많아 발달상권을 기준으로 관측하기로 하였습니다. 각 요일별로 데이터를 관측해 보니 주로 금요일,화요일,목요일,수요일,월요일 같은 평일이 매출이 잘 나오는 것을 관측해 볼 수 있었고 일요일 같은 경우는 매출이 잘 나오지 않음을 관측할 수 있었습니다.

요일이 아닌 시간으로 데이터를 분석해 보자.

시간대별 데이터

	00.06	06.11	11.14	14.17	17.21	21.24
골목상권	3383327	3470377	97310509	83294838	75214374	19398982
관광특구	55973485	1261591943	3393688038	2138814524	1699495163	370793773

발달상권	23452909	403657391	949922721	143496272	586602773	138080096
전통시장	8861086	46710757	151482125	143496272	158737501	52976794



시간대별 데이터를 관측해보니 확실히 점심시간이 포함된 11~14시에 많은 사람들이 이용하며 그 다음으로 14~17시. 17~21시가 비슷하게 사람들이 이용함을 알 수 있었습니다. 그리고 21~24시랑 06~11시는 별로 이용하지 않고 00~06시는 사람들이 거의 이용하지 않음을 알 수 있었습니다. (이 모든 데이터는 24시간 영업해서 매출이 실제로 나오는 카페만을 대상으로 하였다) 종합적인 혼잡도를 계산하기 위해 각 데이터를 병합하고 정규화 과정을 거쳤습니다.

종합 혼잡도(정규화) :

	00.06	06.11	11.14	14.17	17.21	21.00
월요일	0.00425	0.340155	0.822773	0.571646	0.501785	0.105521
화요일	0.00632	0.375784	0.906619	0.630402	0.553562	0.117709
수요일	0.006704	0.382397	0.922181	0.641307	0.563172	0.119971
목요일	0.007207	0.391054	0.942552	0.655583	0.575752	0.122932
금요일	0.008625	0.415465	1	0.695841	0.611227	0.131283
토요일	0.004356	0.341994	0.8271	0.574678	0.504457	0.10615
일요일	0	0.267013	0.650647	0.451025	0.395493	0.080501

다만, 이 데이터는 쉽게 관측하기 어려우므로 여기서 우리는 혼잡도 계산을 편하게 하기 위해 소수점 첫째자리를 기준으로 혼잡도 레벨을 분류하였다. 다음은 혼잡도 레벨별 분류입니다.

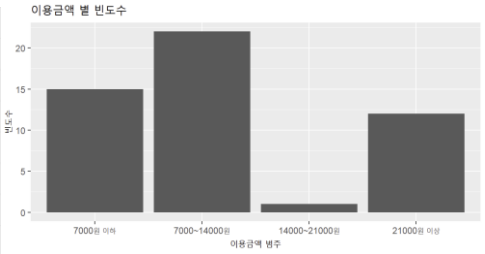
	00~06	6~11	11~14	14~17	17~21	21~00
월요일	0	3	8	5	5	1
화요일	0	3	9	6	5	1
수요일	0	3	9	6	5	1
목요일	0	3	9	6	5	1
금요일	0	4	9	6	6	1
토요일	0	3	8	5	5	1
일요일	0	2	6	4	3	0

결론적으로 각 시간대 그리고 각 요일별 혼잡도에 따른 데이터를 구 할 수 있게 되었습니다. 위 데이터를 통해 얼마나 혼잡한 시간에 카공족들이 들어와 자리를 오랫동안 차지하고 있는지 추론해 낼 계획입니다.

4. 사용자별 이용금액 데이터

여기 부분 부터는 사장님이 공수해오신 자료와 틀을 맞추어야 하기 때문에, 특정한 데이터를 추측하여 만들었다. 셰빌리티 사장님이 구해주신 데이터의 총 50행을 참고하였다. 먼저, 새롭게 만든 첫번째 데이터는 bills[영수증] 데이터다. 사장님의 데이터 collumns의 bill에 해당하는 세부정보를 나타내고 있으며 데이터 구성은 다음과 같다. [bill_id 구매내역, 시간,이용금액]이다. 필수적으로 들어가야할 [장소,user_id]에 관한 데이터가 생략된 이유는 셰빌리티 사장님이 공수해오신 자료와 연관되기 때문에 불필요한 columns을 제거하였다. (참고로 영수증 구매시간이랑 전기 이용시작 시간이랑은 다르다고 판단해 시간 column은 추가로 넣어주었다.)

bill_id	구매내역	이용금액	시간
10	빵 케이크	8000	2023-05-08 15:06
11	음료수 케이크	6000	2023-05-08 15:41
12	빵 케이크	6000	2023-05-08 15:43
13	커피	6000	2023-05-08 15:44
14	커피	5000	2023-05-08 15:45
15	음료수	6000	2023-05-08 15:45
16	음료수	12000	2023-05-08 15:46



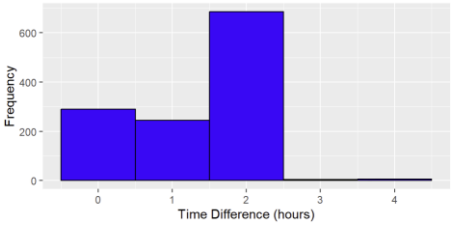
데이터를 분석해본 결과 약 14000원 이내(음료 두잔으로 추정)금액대를 가장 많이 소비하는 것으로 나타났고, 그 다음은 7000원 이내(음료 한잔), 그 다음은 21000원 이상(다인손님 혹은 빵 구매로 추정), 마지막으로 14000~21000원 순으로 카페손님들이 이용금액을 소비한다고 나타났습니다. 이 데이터를 통해 손님의 매출액이 사장님의 만족도에 유의미한 변화를 나타내어 주는지 알아낼 예정이다.

5. 셰빌리티 사장님 데이터 및 각종 생성된 데이터(만족여부,방문회수,이용시간)

셰빌리티 사장님이 공수해오신 자료이다.

name	name	id	user_id	plug	start	request_end	end	bill	status	ticket	create_at
A	1호기	10	1	2	2023-05-08 15:06	NULL	2023-05-08 16:06	10	3	{time":"1"	unit:"hour"
A	1호기	11	3	2	2023-05-08 15:41	NULL	2023-05-08 17:41	11	3	{time":"2"	unit:"hour"
A	1호기	12	3	2	2023-05-08 15:43	NULL	2023-05-08 17:43	12	3	{time":"2"	unit:"hour"
A	1호기	13	1	2	2023-05-08 15:44	NULL	2023-05-08 16:44	13	3	{time":"1"	unit:"hour"
A	1호기	14	1	2	2023-05-08 15:45	NULL	2023-05-08 16:45	14	3	{time":"1"	unit:"hour"
A	1호기	15	3	2	2023-05-08 15:45	NULL	2023-05-08 17:45	15	3	{time":"2"	unit:"hour"
A	1호기	16	8	2	2023-05-08 15:46	NULL	2023-05-08 16:46	16	2	{time":"1"	unit:"hour"
A	1호기	20	13	2	2023-05-08 16:48	NULL	2023-05-08 17:48	20	3	{time":"1"	unit:"hour"
A	1호기	22	3	2	2023-05-08 19:05	NULL	2023-05-08 21:05	22	2	{time":"2"	unit:"hour"
A	1호기	23	3	2	2023-05-09 10:01	NULL	2023-05-09 11:01	23	3	{time":"1"	unit:"hour"
A	1호기	25	3	2	2023-05-09 10:35	NULL	2023-05-09 11:35	25	3	{time":"1"	unit:"hour"

먼저 간단히 데이터 분석부터 해보자. End-start 시간을 계산해서 이용자들이 얼마나 오래 카페전기를 사용했는지 알아보자.



대부분이 데이터가 2시간 미만으로 측정되어 있음을 알 수 있다. 이는 오랜시간 카페에 머무는 카공족들의 데이터를 충분히 반영하지 못하는 것 같아 아쉬움이 있다.

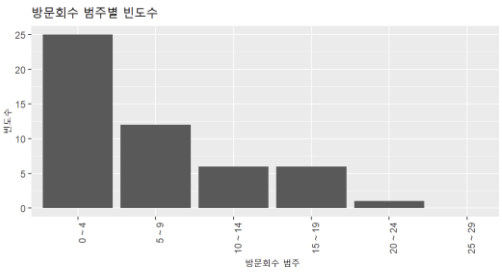
이 데이터를 통해 특정 사람이 몇시간동안 플러그 몇 개를 사용해서 전기를 사용했는지 알 수 있는 데이터이다. 이 데이터를 통해 전기를 각 사람이 얼마나 많이 사용했는지 뽑아낼 계획이다. 그리고, 이용 시간대가 나와 있으므로 위에 혼잡도 데이터랑 연결하여 얼마나 혼잡할 때 왔는지 또한 뽑아낼 계획이다.

다만, 종합적인 결론을 내리기에 앞서 아래 그림과 같이 user_id의 1인 사람이 5월10일 4시25분에 1시간 전기를 사용했고 또 같은날 4시 26분에 동시에 1시간 전기를 사용했고 또 같은날 오후 5시 19분에 1시간 전기를 사용했고.. 같이 시간상 일어날 수 없는 일이 종종 일어났다. 이는 데이터의 신뢰성에 악영향을 줄 수 있으므로 user_id와 가게명이 같은 경우 하나의 데이터만 남기고 나머진 삭제 하였다. 그 후 데이터 분석을 위해 상위 50개의 데이터만 남겼다.

B	1호기	49	1	3	2023-05-10 16:25 NULL	2023-05-10 17:25	49	3 {time":"1"
B	2호기	50	1	4	2023-05-10 16:26 NULL	2023-05-10 17:26	50	3 {time":"1"
B	3호기	52	1	5	2023-05-10 17:19 NULL	2023-05-10 18:19	52	3 {time":"1"
B	2호기	53	1	4	2023-05-10 17:20 NULL	2023-05-10 18:20	53	3 {time":"1"
B	3호기	54	1	5	2023-05-10 17:21 NULL	2023-05-10 18:21	54	3 {time":"1"
B	1호기	55	1	3	2023-05-10 17:40 NULL	2023-05-10 18:40	55	3 {time":"1"

그 다음은 특정 user가 해당 카페(서비스 하는 카페)를 한달동안 얼마나 자주 방문했는가에 대한 자료이다.(카페에서 전기를 사용하지 않더라도 방문은 했을 수 있으므로 카페사장님의 자료에서 따오지 않고 새로 만들었다.)

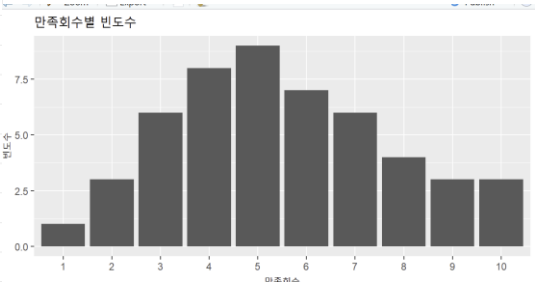
user_id	카페명	방문회수
1	A	5
1	B	2
1	C	3
3	A	8
8	A	7



데이터 분석 결과 역시 0~4회 방문하는 단골이 아닌 고객들이 가장 많고 방문횟수의 범주가 높아질 수록, 빈도수도 적어지는 결과가 나타났다.

그 다음은 카페 사장님 만족도 자료이다. 특정 user에 따른 카페 사장님들이 얼마나 만족하는지 결과를 내포하고 있다. 위 데이터는 특정 user에 대한 카페 사장님의 만족도를 평가한다.

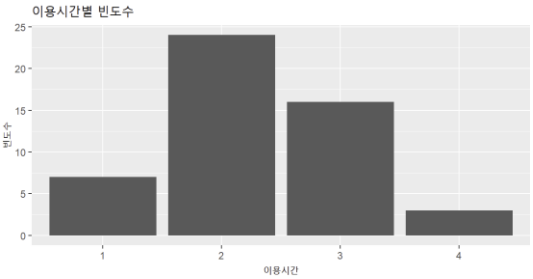
user_id	카페명	만족여부
1	A	0
1	B	0
1	C	1
3	A	1
8	A	1
11	A	2
13	A	2
16	A	2
17	A	2
20	B	2



만족도는 정규분포를 완벽히 따르고 있지는 않지만 나름 중앙값인 4~6 사이가 높게 분포하고 0과 10쪽으로 갈수록 값이 빈도수가 줄어드는 모습을 관찰할 수 있었다.

그 다음은 이용시간 데이터이다.

카페명	user_id	이용시간
A	1	2
B	1	2
C	1	2
A	3	2
A	8	2
A	11	2



이용시간은 2시간이 가장 많았고 그다음은 3시간이 가장 많았다 4시간 이상 이용하는 장기 고객은 적은 것으로 나타났다.

6. 데이터 병합

이제 종합적인 결론을 내리기 위해 각 데이터를 필요한 칼럼을 불러와 결합해 보겠다. 결합 방식은 다음과 같다.

1. data 데이터에서 필요한 column만 남기기 (name...1, name...2, user_id, plug, start, end, bill)
2. 새로운 데이터 구조를 만들기 (이름 = merged_data)
3. merged_data에 data 데이터를 불러오고 bill 데이터랑 결합하기 (data데이터의 51번째 줄까지만 불러와서 bill 데이터랑 결합)
4. merged_data의 user_id, name...1의 조합이 satisfy 데이터의 user_id, 카페명과 같을때의 satisfy데이터의 만족여부 column에 해당하는 값을 가져와 merged_data에 열을 추가하기
5. 4번에 진행했던것과 마찬가지로, 이번엔 merged_data의 user_id, name...1의 조합이 vip 데이터의 user_id, 카페명과 같을때의 방문회수 column에 해당하는 값을 가져와 merged_data에 열을 추가하기
6. merged_data의 시간 column에 해당하는 값의 요일과 시간대를 구해서 혼잡도에 있는 표를보고 해당하는 값을 가져와 merged_data에 혼잡도 라는 열을 추가하기
7. plug*(end-start)로 전기 사용량 계산
8. 실제로 t-test 및 분석에서 사용하는 column만 남기기(name...1, user_id, bill, 이용금액, 만족여부, 방문회수, 혼잡도, 전기사용량)

이 모든 과정을 거친 후 남겨진 데이터는 다음과 같은 모양이다.

name...1	user_id	bill	이용금액	만족여부	방문회수	혼잡도	전기사용량
B	23	69	23000	3	17	6	8
A	3	68	20000	1	8	6	2
B	25	60	16000	3	15	0	8
B	24	59	10000	3	2	0	10
C	26	65	17000	3	14	6	18
C	1	61	36000	1	3	6	7
C	1	63	19000	1	3	6	9
C	1	64	14000	1	3	6	9
C	1	62	12000	1	3	6	8
C	27	66	20000	4	11	5	18
B	28	67	24000	4	5	1	5

7. T-test 및 regression 시작

먼저 이용금액에 따른 t-test결과이다. 알 수 있게 되었다. 이용금액 1만원을 기준으로 분류한 후 만족여부의 차이에 따른 t-test를 수행해 보았다. 그 결과, P-value가 1.364e-07로써, 0.05보다 작으므로 '이용금액(1만원보다 큰지 작은지)에 따라 카페사장님들의 만족도가 차이가 있다'가 통계적으로 유의미한 결론임을 알 수 있었다.

```
data:  high_spend_group and low_spend_group
t = 7.4235, df = 23.376, p-value = 1.364e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.711953 4.804714
sample estimates:
mean of x mean of y
 7.933333  4.175000
```

그 다음은 혼잡도에 따른 t-test결과이다. 혼잡도 4를 기준으로 분류한 후 만족여부의 차이에 따른 t-test를 수행해 보았다. 그 결과, P-value가 0.000118로써, 0.05보다 작으므로 '혼잡도(4보다 큰지 작은지)에 따라 카페사장님들의 만족도가 차이가 있다'가 통계적으로 유의미한 결론임을 알 수 있게 되었다.

```
        Welch Two Sample t-test

data:  high_group and low_group
t = -4.2806, df = 38.829, p-value = 0.000118
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.586804 -1.284625
sample estimates:
mean of x mean of y
 4.314286  6.750000
```

그 다음은 방문횟수(단골여부)에 따른 t-test결과이다. 한달에 15번 이상 방문하면 단골이라고 추측하여 방문횟수 15를 기준으로 t-test를 수행해 보았다. 그 결과, P-value가 0.2585로써, 0.05보다 크므로 '방문횟수(15보다 큰지 작은지)에 따라 카페사장님들의 만족도가 차이가 있다'가 통계적으로 유의미하지 않은 결론임을 알 수 있게 되었다.

```
        Welch Two Sample t-test

data:  high_group and low_group
t = -1.1741, df = 15.131, p-value = 0.2585
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.407518  0.696407
sample estimates:
mean of x mean of y
 4.500000  5.355556
```

그 다음은 전기사용량 에 따른 t-test결과이다. 4시간동안 2개이상의 플러그를 사용할 경우 많은 전기를 사용한다고 추측하여, 전기사용량 8(이용시간*플러그 수)를 기준으로 t-test를 수행해 보았다. 그 결과, P-value가 0.8155로써, 0.05보다 크므로 '전기사용량 (8보다 큰지 작은지)에 따라 카페 사장님들의 만족도가 차이가 있다'가 통계적으로 유의미하지 않은 결론임을 알 수 있게 되었다. 이는 모든 데이터의 전기 사용 시간이 1~2시간으로 측정되어졌기 때문이라고 추측된다.

```
        Welch Two Sample t-test

data:  high_group and low_group
t = -0.23627, df = 21.478, p-value = 0.8155
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.794838  1.428172
sample estimates:
mean of x mean of y
 5.066667  5.250000
```

종합적으로 카페사장님들의 만족도에 통계적으로 영향을 끼치는 유의미한 변수는 시간별 혼잡도와 이용금액으로 나타났다.

그 다음 유의미한 변수들을 대상으로 회귀분석을 수행해 보았다.

```
Call:
lm(formula = 만족여부 ~ 혼잡도 + 이용금액, data = merged_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.92201 -0.75521  0.01146  0.90499  2.63633

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.585e+00  3.809e-01  14.661  < 2e-16 ***
혼잡도      -3.902e-01  7.430e-02  -5.252 2.83e-06 ***
이용금액     5.767e-05  6.553e-06   8.800 7.10e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.218 on 52 degrees of freedom
  (결측으로 인하여 1개의 관측치가 삭제되었습니다.)
Multiple R-squared:  0.7346,    Adjusted R-squared:  0.7244
F-statistic: 71.96 on 2 and 52 DF,  p-value: 1.051e-15
```

P-value의 값은 전부 0.05보다 한참 작으므로 통계적으로 각 변수들의 값이 유의미하라는걸 알 수 있으며 R-squared값과 Adjusted R-squared 값 또한 각각 0.7346, 0.7244로 회귀식으로 나온 방정식이 유의미하다라는 것을 알 수 있다. 종합적으로, 만족도 = -3.902e-01*혼잡도 + 5.767e-05*이

용금액 + 5.585e+00라는 식을 도출해 낼 수 있었다.

그다음은 로지스틱 회귀분석을 실시 해 보았다.

로지스틱 회귀분석을 위해 만족도를 4를 기준으로 이진 분류 하였다. 그 후 혼잡도와 이용금액에 따른 로지스틱 회귀분석을 돌렸다. 하지만 전체적으로 단순회귀모형보다 p-value가 떨어짐을 관측할 수 있다. 이는 만족도가 이진으로 나타낼 수 있는 분류가 아니라 1~10까지의 연속적인 값을 가지기 때문이라고 추측할 수 있다.

```
Call:
glm(formula = binary_satisfaction ~ 혼잡도 + 이용금액,
    family = binomial(link = "logit"), data = merged_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.5787574  3.0274041  -0.191   0.8484
혼잡도       -0.9832854  0.4850774  -2.027   0.0427 *
이용금액      0.0008801  0.0003552   2.477   0.0132 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.021  on 54  degrees of freedom
Residual deviance: 34.849  on 52  degrees of freedom
  (결측으로 인하여 1개의 관측치가 삭제되었습니다.)
AIC: 40.849

Number of Fisher Scoring iterations: 9
```

따라서 처음에 사용한 회귀식을 사용하기로 결정했다. (만족도 = -3.902e-01*혼잡도 + 5.767e-05*이용금액 + 5.585e+00)

8. 결론

플러그를 사용함으로써 고객들에게 돈을 받기 위해서는 카페사장님들의 만족도에 따라 다르게 돈을 받는 것이 적합하다고 생각이 든다. 각 항목을 반올림 하여 round((10-(-0.4*혼잡도 + 0.57*이용금액(단위 : 1000) + 5.585))*500,3)(단위 : 원)(최소 500원, 최대 3500원) 만큼의 이용금액을 2시간 마다 받는 것(데이터 상의 전기 이용시간이 최대 2시간이였기 때문에)을 추천한다. 예를들어

1. 수요일 3시(6의혼잡도)에 방문해서 10000원의 비용을 낸 사람에게는 2시간에 600원의 이용금액
2. 금요일 2시(9의 혼잡도)에 방문해서 5000원의 비용을 낸 사람에게는 2시간에 3400원의 이용금액
3. 월요일 새벽3시(0의 혼잡도)에 방문해서 50000원의 비용을 낸 사람에게는 공짜로 전기를 사용할 수 있게 해주는 것이다.

이는 우리가 통상적으로 생각하는 기준과 올바르게 떨어진다. 혼잡한 시간에 오더라도 2시간에 20000원이상 금액을 사용한 사람에게는 전기를 사용할 수 있게 하고, 만약 굉장히 혼잡한 시간에

커피한잔 시키고 전기를 사용할려는 사람에게는 2시간에 약 3500원 정도의 요금이 부과 되며, 굉장히 한가한 시간에 온다면 커피한잔을 시키더라도 2시간에 1000원 정도의 적당한 요금만이 부과되기 때문이다.

9. 사용한 데이터 설명

congestion.R : 서울시 상권분석 서비스 데이터에서 요일별, 그리고 시간대별 혼잡도를 뽑아내기 위해 groupby를 사용한 R 코드입니다.

network.R : 네트워크분석, wordcloud분석 등을 수행하기 위해 사용한 R 코드입니다.

t-test.R : 각 데이터들을 병합하고 t-test, 선형회귀분석, 로지스틱 회귀분석 등을 수행한 코드입니다. 또한, 마지막에 각종 그래프들을 뽑아내기 위한 코드 또한 담겨져 있습니다.

네이버뉴스_카공족.xlsx : 네이버 뉴스에서 "카공족" 이라는 키워드로 검색하여 크롤링한 결과입니다. 이는 network.R 파일에서 네트워크분석 및 워드클라우드 분석을 사용하는데 쓰였습니다.

데이터생성.xlsx : 각종 제가 직접 만든 데이터들입니다. 이 데이터 내부에는 카페 이용시간, 카페 혼잡도, 카페 영수증, 카페방문회수, 카페사장님 만족도 데이터 등이 있습니다.

사용데이터 (1).xlsx : 셰빌리티 사장님께서 전해주신 카페별 전기 사용ㄹ야 데이터 입니다.

서울시 상권분석서비스(추정매출-상권).csv : 서울시 상권분석 데이터입니다. congestion.R 코드를 통해 카페 영업하시는 분들만 따로 뽑아내고 또한 groupby를 통해 각 상권에 따른 요일별, 시간별 매출액으로 데이터생성.xlsx에 있는 카페 혼잡도 데이터를 뽑아내었습니다.

카페사장님_자유게시판.xlsx : 인증된 카페사장님들만 쓸 수 있는 자유게시판 제목 키워드를 크롤링한 결가ㅗ입니다. 이는 network.R 분석에서 네트워크분석을 사용하는데 쓰였습니다.