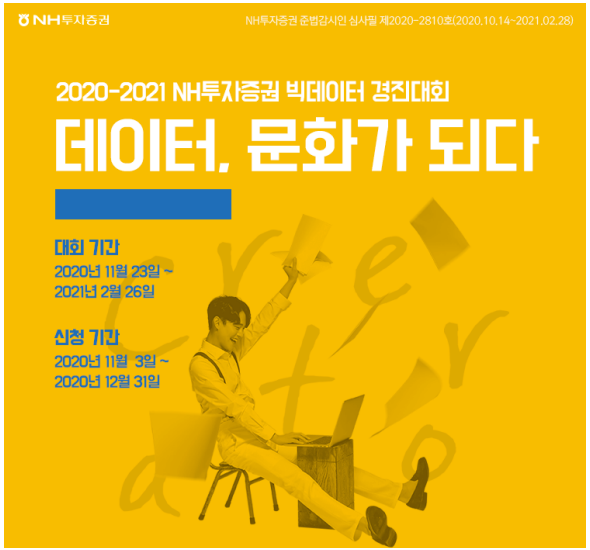
Investor Profiling

문구영 성유지 이영신

KU-BIG

NH 투자증권 빅데이터 분석대회



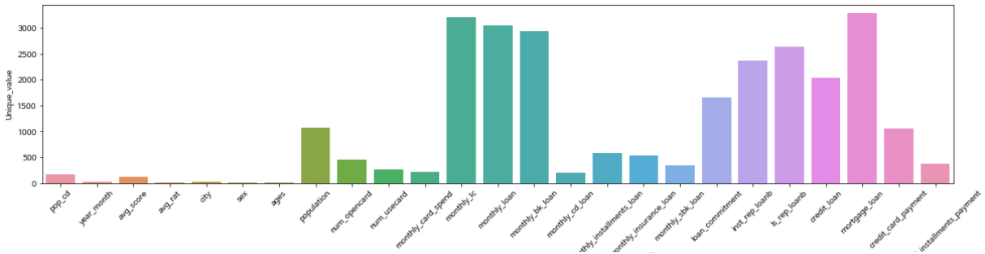
- League 2: Y&Z 세대 투자자 프로파일링
- 1980년대 후반부터 현재까지의 출생연도를 가진 사람들의 투자 행동 분석
- 주어진 데이터로 다양한 인사이트를 끌어내야
- 시각화, 데이터 마이닝

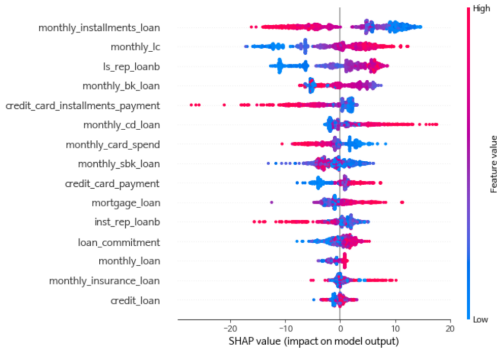
저희 조의 계획

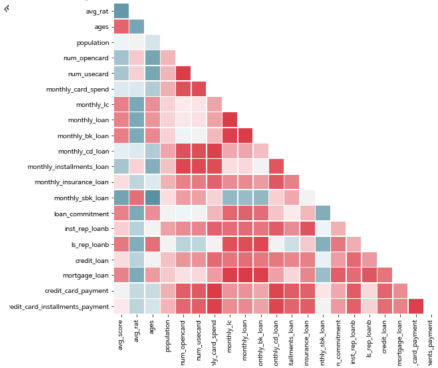

2020 11 NOV
일 SUN
월 MON
화 TUE
수 WED
목 THU
금 FRI
토 SAT
유사 주제의 데이터 분석 대회 수상작 코드 구현
금융 분야에서 활용되는 빅데이터 기법 공부
주어진 데이터 EDA

유사 주제의 데이터 분석 대회 수상작 코드 구현

KCB 금융스타일 시각화 경진대회





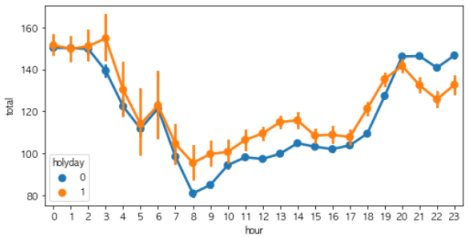


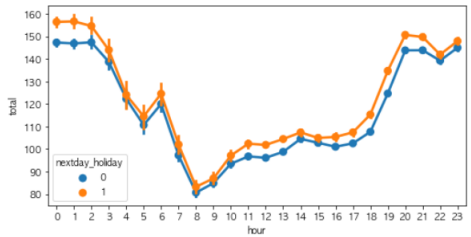
1. 데이터 시계열 처리
2. 변수가 취할 수 있는 유니크한 값들 (변수 값의 다양성)
3. 칼럼간 상관관계 시각화, 정렬 (unstack 이용)
4. 변수 별 상관관계 (groupby)

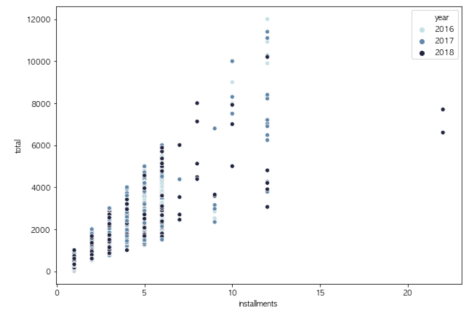
* 타깃 변수에 영향을 많이 주는 요소 분석 (예측 모델 사용)
5. lgbm, xgboost 변수 중요도 분석 - 변수가 얼마나 영향을 주는 지
6. shap 중요도 분석 - 변수가 영향을 끼치는 방향 (높이는 지, 낮추는 지)

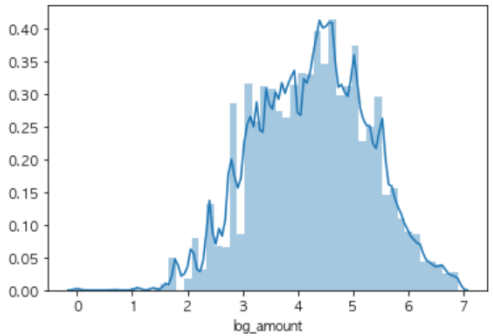
유사 주제의 데이터 분석 대회 수상작 코드 구현

신용카드 거래 데이터 시각화







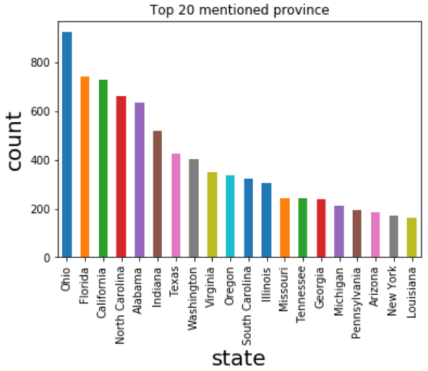


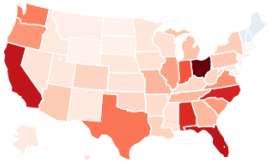
1. 특이값 탐색 (catplot, np.percentile 이용)
2. 변수간 상관관계 (heatmap)
3. 변수가 취할 수 있는 값들의 빈도수 (countplot)
 -> 빈도수 상위 n개 추출
4. 각 기준(매출액, 거래기록)에 따른 상위 n개의 store_id (.loc)
5. 세분화된 시계열 데이터 생성 (pd.datetime)
 -> Y / M / D / H /M / S 열 생성
 -> 연도별, 월별, 시간별 데이터

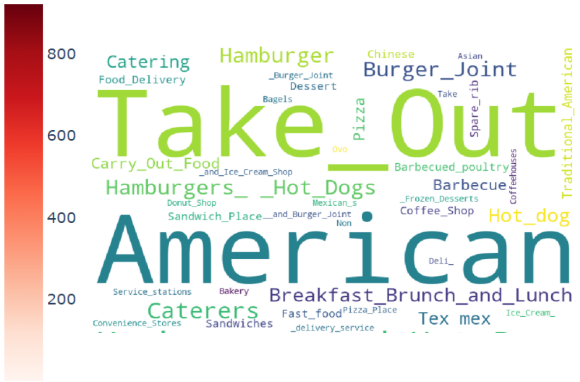
* 다음 날이 주말인지, 공휴일인지 여부를 새로운 변수로 만들어서 시각화

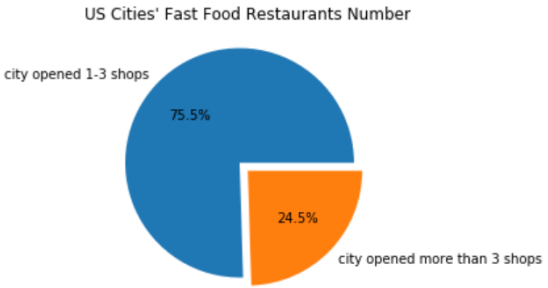
유사 주제의 데이터 분석 대회 수상작 코드 구현

Kaggle









교수님 면담 및 자문

김중혁 교수님의 자문

- 20대, 30대와 다른 연령대와의 비교
- 20대, 30대 내에서 다양한 기준을 가지고 분류
  (ex. 자산, 인당 계좌 수)
- Behavior Finance에 대한 공부

※ 데이터 셋이 나오기 전에 면담 진행.
추후에 교수님과의 면담 재진행 예정.

주어진 데이터 분석

CUS_INFO.CSV(10,000건)
ACT_INFO.CSV(23,959건)
IEM_INFO.CSV(5,065건)
TRD_KR.CSV(3,312,664건)
TRD_OSS.CSV(29,301건)
data_schema_vf: 데이터 명세

CUS_INFO.CSV(10,000건)

No	컬럼명	컬럼한글명	컬럼설명	
1	CUS_ID	고객번호	고객을 구분할 수 있는 Unique값	1 b 2 8
2	SEX_DIT_CD	성별	1: 남성 / 2: 여성	1
3	CUS_AGE	연령대	00: 19세 이하 20: 20~24세 / 25: 25~29세 30: 30~34세 / 35: 35~39세 40: 40~44세 / 45: 45~49세 50: 50~54세 / 55: 55~59세 60: 60~64세 / 65: 65~70세 70: 70세 이상	1(
4	ZIP_CTP_CD	주소(시도)	41: 경기 / 11: 서울 / 48: 경남 / 26: 부산 / 27: 대구 47: 경북 / 28: 인천 / 44: 충남 / 46: 전남 / 30: 대전 29: 광주 / 43: 충북 / 45: 전북 / 42: 강원 / 31: 울산 50: 제주 / 36: 세종	4
5	TCO_CUS_GRD_CD	고객등급	01: 탑클래스 (자산[1]10억이상 or 수익기여도[2] 5백만원 이상) 02: 골드 (자산3억이상 or 수익기여도 3백만원 이상) 03: 로얄 (자산1억이상 or 수익기여도 1백만원 이상) 04: 그린 (자산3천이상 or 수익기여도 5십만원 이상) 05: 블루 (자산1천이상 or 수익기여도 1십만원 이상)	5
6	IVS_ICN_CD	고객투자성향	01: 안정형 / 02: 안정추구형 / 03: 위험중립형 04: 적극투자형 / 05: 공격투자형 09: 전문투자자형 / 00:정보제공미동의	4

주어진 데이터 분석

CUS_INFO.CSV(10,000건)
ACT_INFO.CSV(23,959건)
IEM_INFO.CSV(5,065건)
TRD_KR.CSV(3,312,664건)
TRD_OSS.CSV(29,301건)
data_schema_vf: 데이터 명세

IEM_INFO.CSV(4,685건)

No	컬럼명	컬럼한글명	컬럼설명	
1	IEM_CD	종목코드	주식종목코드	A
2	IEM_ENG_NM	종목영문명	종목 영문명	D
3	IEM_KRL_NM	종목한글명	종목 한글명	동

TRD_KR.CSV(3,312,664건)

No	컬럼명	컬럼한글명	컬럼설명	
1	ACT_ID	계좌번호	계좌를 구분할 수 있는 Unique값	7 a f b
2	ORR_DT	주문날짜	주문날짜	2
3	ORR_ORD	주문순서	주문날짜 기준으로 주문 순서	3
4	ORR_RTN_HUR	주문접수시간대	주문접수시간대	9
5	LST_CNS_HUR	최종체결시간대	최종체결시간대	1
6	IEM_CD	종목코드	거래소 종목코드	A
7	SBY_DIT_CD	매매구분코드	1: 매도 / 2: 매수	1
8	CNS_QTY	체결수량	종목 체결수량	5
9	ORR_PR	체결가격	종목 체결원화단가	1
10	ORR_MDI_DIT_CD	주문매체구분코드	거래소 주문전달용 매체구분 0: 영업점단말 / 1: 유선단말 / 2: 무선단말MTS / 3: HTS / 4: 기타	3

주어진 데이터 분석

CUS_INFO.CSV(10,000건)
ACT_INFO.CSV(23,959건)
IEM_INFO.CSV(5,065건)
TRD_KR.CSV(3,312,664건)
TRD_OSS.CSV(29,301건)
data_schema_vf: 데이터 명세

CUS_INFO.CSV(10,000건)

No	컬럼명	컬럼한글명	컬럼설명	
1	CUS_ID	고객번호	고객을 구분할 수 있는 Unique값	1 b 2 8
2	SEX_DIT_CD	성별	1: 남성 / 2: 여성	1
3	CUS_AGE	연령대	00: 19세 이하 20: 20~24세 / 25: 25~29세 30: 30~34세 / 35: 35~39세 40: 40~44세 / 45: 45~49세 50: 50~54세 / 55: 55~59세 60: 60~64세 / 65: 65~70세 70: 70세 이상	1(
4	ZIP_CTP_CD	주소(시도)	41: 경기 / 11: 서울 / 48: 경남 / 26: 부산 / 27: 대구 47: 경북 / 28: 인천 / 44: 충남 / 46: 전남 / 30: 대전 29: 광주 / 43: 충북 / 45: 전북 / 42: 강원 / 31: 울산 50: 제주 / 36: 세종	4
5	TCO_CUS_GRD_CD	고객등급	01: 탑클래스 (자산[1]10억이상 or 수익기여도[2] 5백만원 이상) 02: 골드 (자산3억이상 or 수익기여도 3백만원 이상) 03: 로얄 (자산1억이상 or 수익기여도 1백만원 이상) 04: 그린 (자산3천이상 or 수익기여도 5십만원 이상) 05: 블루 (자산1천이상 or 수익기여도 1십만원 이상)	5
6	IVS_ICN_CD	고객투자성향	01: 안정형 / 02: 안정추구형 / 03: 위험중립형 04: 적극투자형 / 05: 공격투자형 09: 전문투자자형 / 00:정보제공미동의	4

주어진 데이터 분석

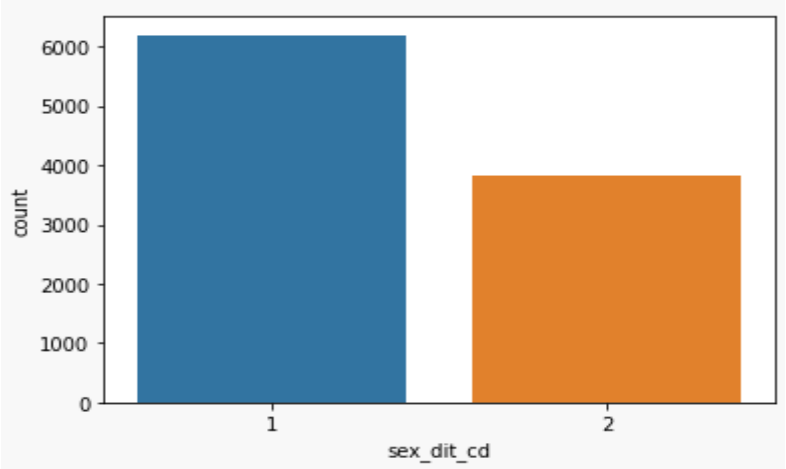
CUS_INFO.CSV(10,000건)
ACT_INFO.CSV(23,959건)
IEM_INFO.CSV(5,065건)
TRD_KR.CSV(3,312,664건)
TRD_OSS.CSV(29,301건)
data_schema_vf: 데이터 명세

TRD_OSS.CSV(29,301건)

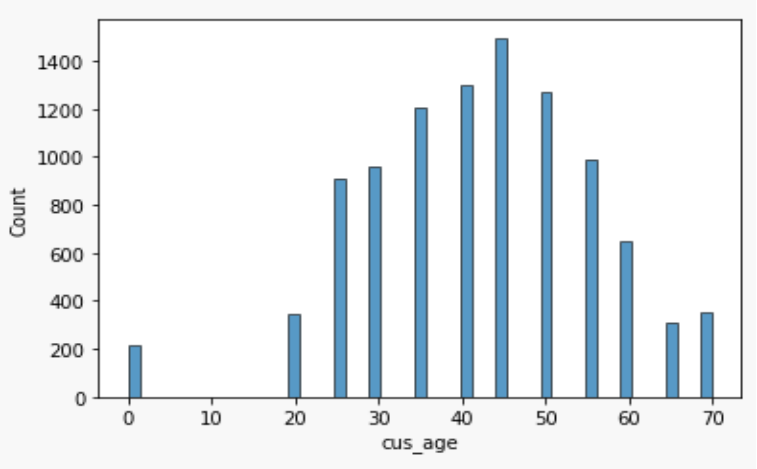
No	컬럼명	컬럼한글명	컬럼설명	
1	ACT_ID	계좌번호	계좌를 구분할 수 있는 Unique값	7 a f b
2	ORR_DT	주문날짜	주문날짜	2
3	ORR_ORD	주문순서	주문날짜 기준으로 주문 순서	3
4	ORR_RTN_HUR	주문접수시간대	주문접수시간대	9
5	LST_CNS_HUR	최종체결시간대	최종체결시간대	1
6	IEM_CD	종목코드	거래소 종목코드	C
7	SBY_DIT_CD	매매구분코드	1: 매도 / 2: 매수	1
8	CNS_QTY	체결수량	종목 체결수량	5
9	ORR_PR	체결가격	종목 체결외화단가	1
10	ORR_MDI_DIT_CD	주문매체구분코드	거래소 주문전달용 매체구분 0: 영업점단말 / 1: 유선단말 / 2: 무선단말 / 3: HTS / 4: 기타	3
11	CUR_CD	거래통화코드	AUD: 오스트레일리아-달러 / CAD: 캐나다-달러 / CNY: 렌민비(위안) EUR: 유로 / GBP: 영국-파운드/ HKD: 홍콩-달러 / IDR: 루피아 JPY:일본-엔 / KRW: 대한민국-원 / SGD: 싱가포르-달러 USD: 미국-달러 / VND: 동	A

주어진 데이터 분석

간단한 시각화 구현



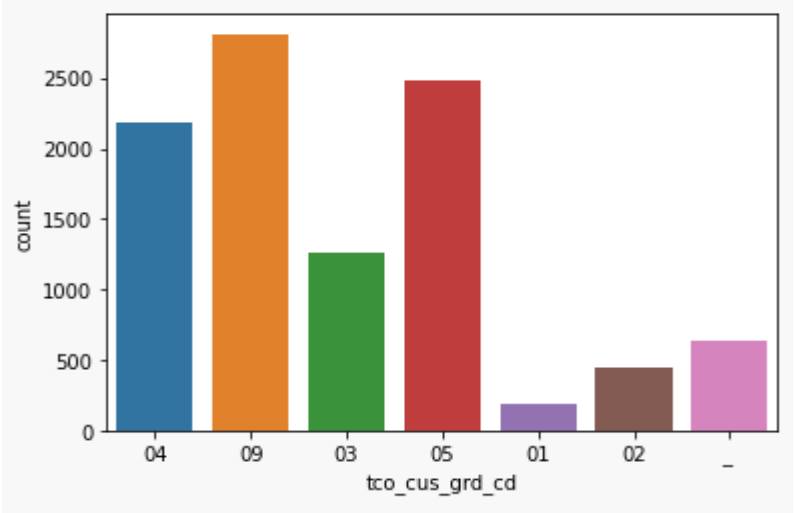
남성 > 여성



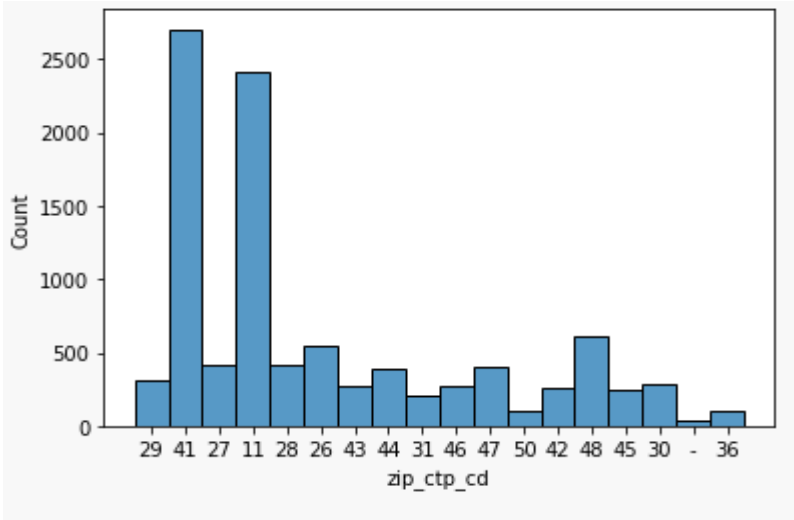
연령대 별 분포

주어진 데이터 분석

간단한 시각화 구현



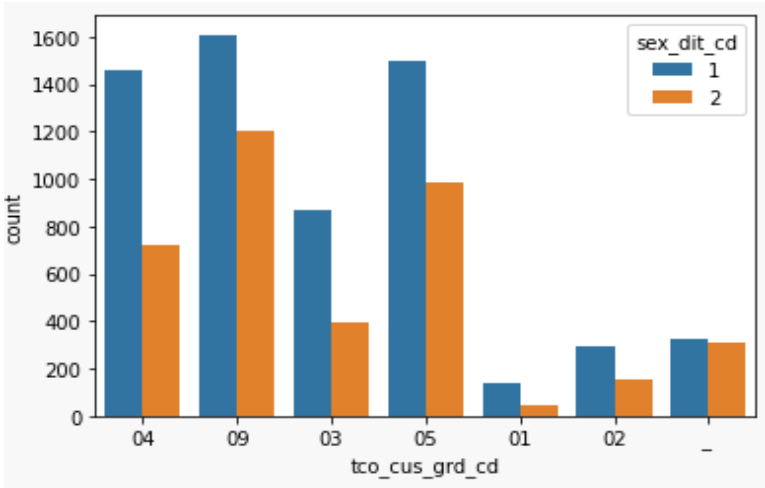
고객 등급 별 분포



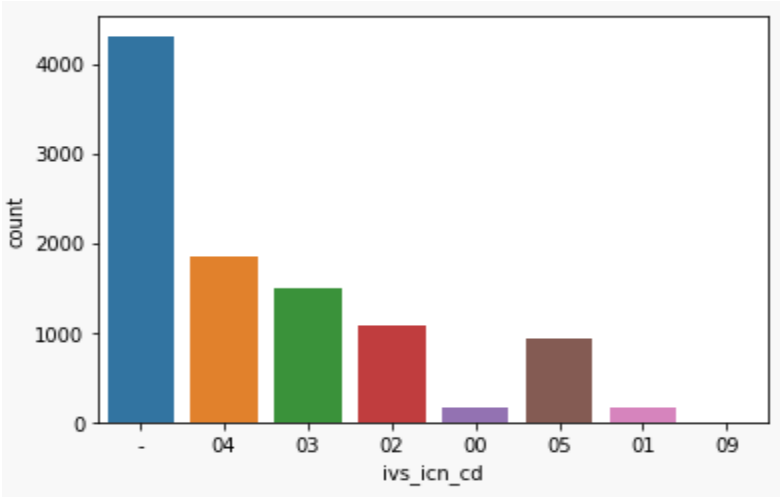
지역 별 분포

주어진 데이터 분석

간단한 시각화 구현



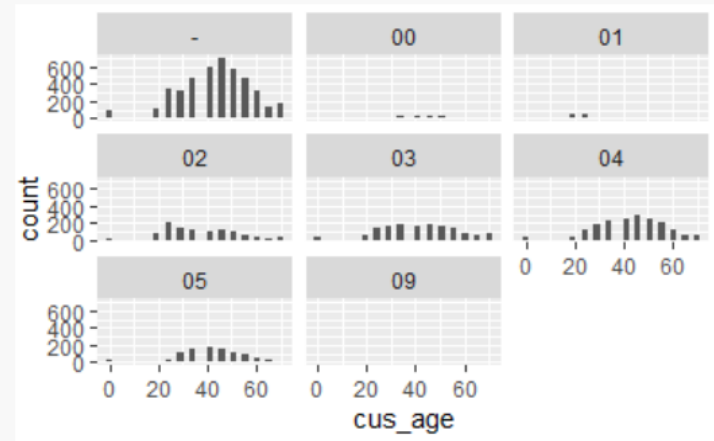
성별에 따른 고객 투자 성향



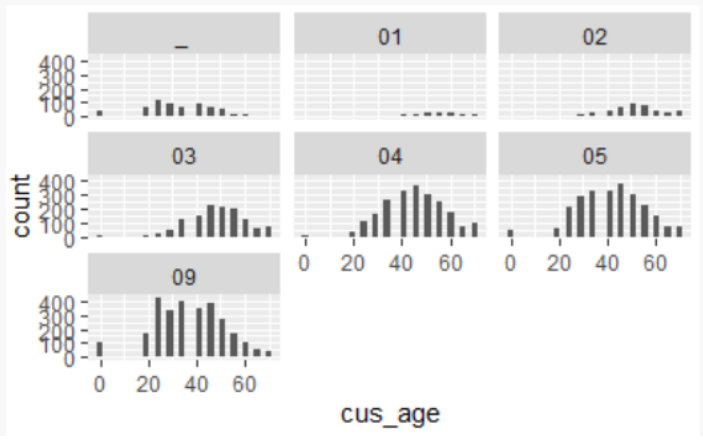
고객 투자 성향 별 분포

주어진 데이터 분석

간단한 시각화 구현



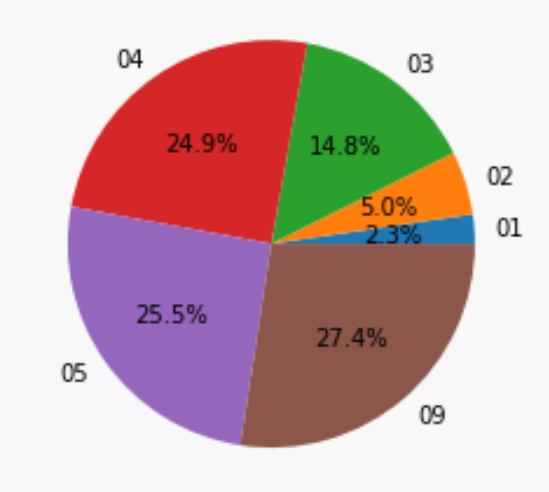
고객 투자성향 별 연령대 분포



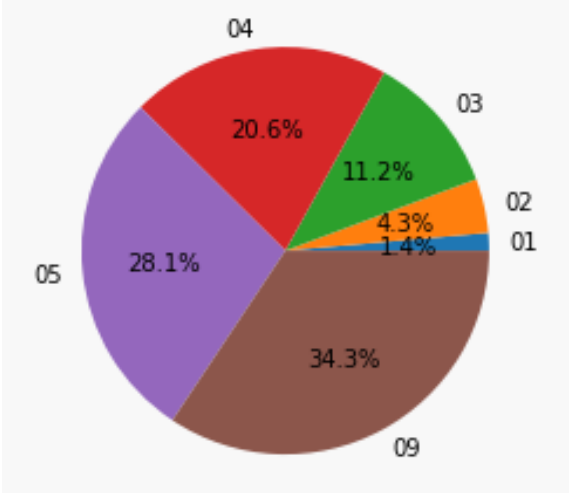
고객 등급 별 연령대 분포

주어진 데이터 분석

간단한 시각화 구현



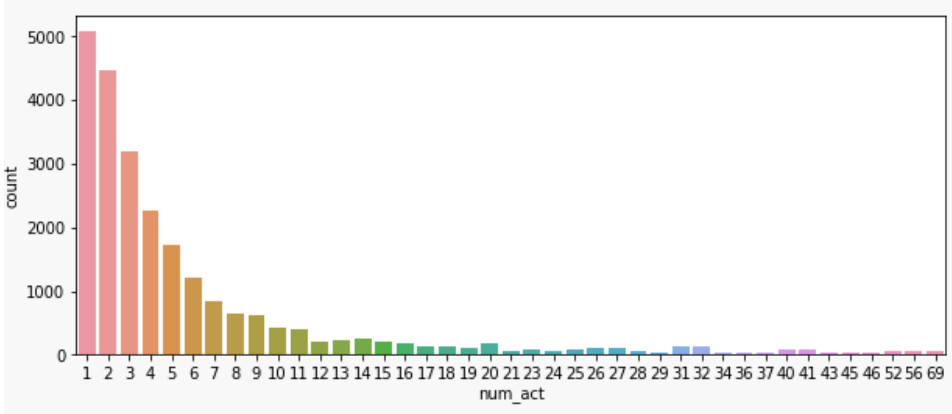
고객 등급 파이차트(남성)



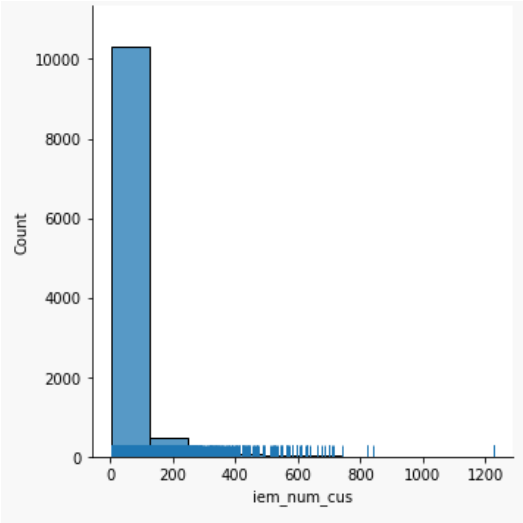
고객 등급 파이차트(여성)

주어진 데이터 분석

간단한 시각화 구현



인 당 계좌 수



계좌 별 종목 개수

주어진 데이터 분석

논의 사항

-데이터 합병의 중요성
고객번호를 기준으로 고객 관련 자료에 계좌 관련 자료 합침.

-체결 수량과 가격을 곱한 금액 계산

-평균적으로 얼마나 수익을 냈는지

-매도와 매수 간의 시간 차이를 통해 투자 성향 파악(단기투자, 장기투자)

-종목 개수를 통해 투자 성향 파악(분산투자, 집중투자)

-2030을 중심으로 분석(2030내 패턴, 2030과 다른 연령대와의 비교)
계좌 총 10997개 중 3696개가 2030
총 9909명 중 3374명이 2030

주어진 데이터 분석

cus_id	sex	age	home	grade	age_cat	num_act	iem_num	iem_num_cus	orr_dt	orr_ord	ord	real	iem_cd	buysell	amount
82f5698372aea023bd...	1	65	11	01	60	40	2	2	20200206	1	10	10	A001440	2	6170000.0
82f5698372aea023bd...	1	65	11	01	60	40	2	2	20200219	1	10	10	A001440	2	10782000.0
82f5698372aea023bd...	1	65	11	01	60	40	2	2	20200218	1	10	10	A001440	2	1210000.0
82f5698372aea023bd...	1	65	11	01	60	40	2	2	20200623	1	10	10	A010140	1	97362790.0
41cb14e8e75685278a...	1	35	28	09	30	3	5	5	20200320	1	11	11	A005930	2	449000.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
73cadd9f69b6a33b4b...	2	35	27	09	30	1	4	4	20200519	2	12	13	A016100	2	19000.0
73cadd9f69b6a33b4b...	2	35	27	09	30	1	4	4	20200526	1	10	10	A187790	2	16000.0
73cadd9f69b6a33b4b...	2	35	27	09	30	1	4	4	20200514	1	14	15	A005930	2	47750.0
73cadd9f69b6a33b4b...	2	35	27	09	30	1	4	4	20200514	2	14	14	A005930	2	47800.0
505afd467e11879571f...	1	35	48	09	30	1	1	1	20190408	1	9	9	A003000	2	127400.0

데이터 합병 및 변수 추가 한 데이터셋

주어진 데이터 분석

논의 사항

6개의 데이터 셋 구축

1. 종목 개수
2. 매도/매수 총 금액 -> 투자의 크기 가늠
3. 매수 후 매도가 이루어진 거래에 한해서 :
   거래 시간차 , 거래 수익(손실)
4. 총 수익
5. 자산 규모(고객 등급으로 가늠) 대비 매도/매수 금액
6. 매수 금액 대비 수익

감사합니다!☺

KU-BIG