제 2회 신빅해

# 신한 라이프

블루 라이프(숙명여대 / 강예원, 김예림, 김예지)

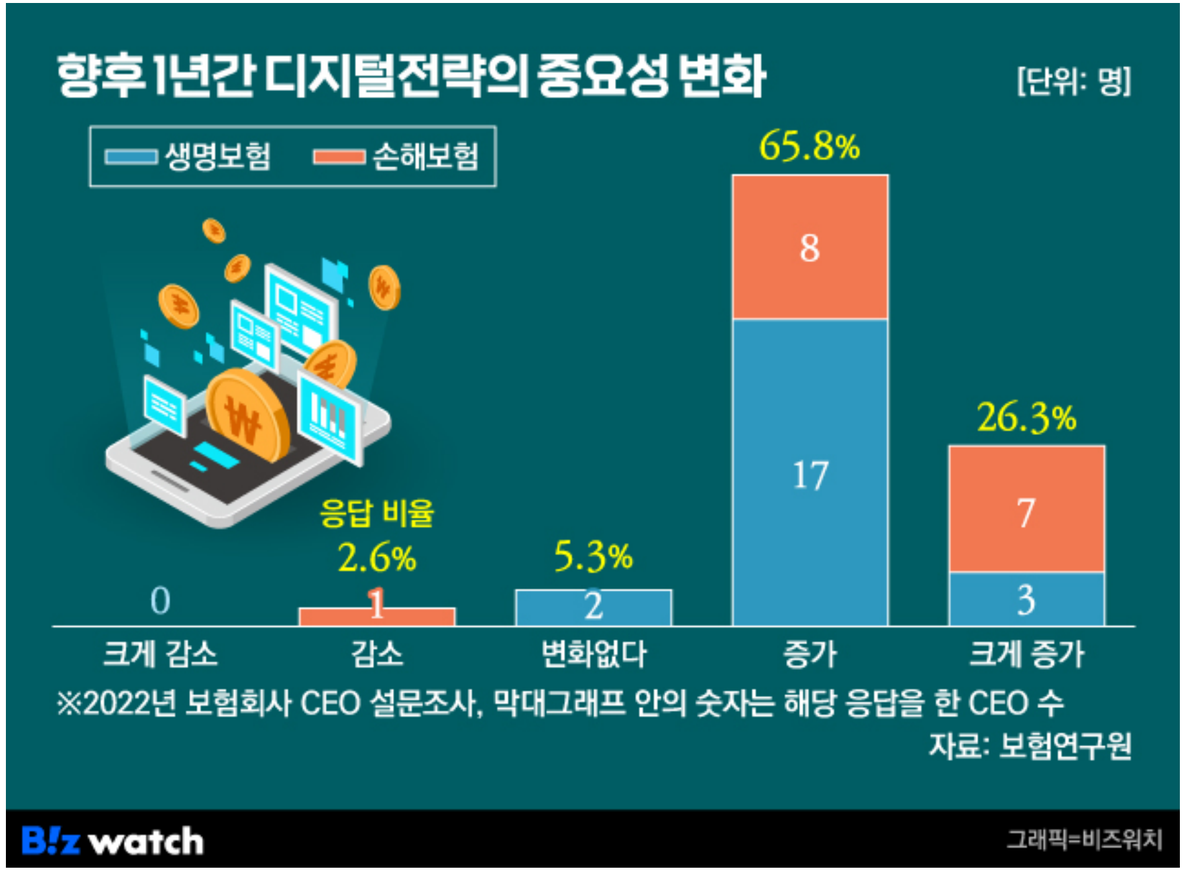
# CONTENTS

제2회

# 신박해

1. 개요
2. 데이터 전처리
3. 데이터 분석
4. 데이터 모델링 및 예측
5. 신사업 [청년 연금저축 보험]

# 1.개요




- 코로나19 이후 금융소비자들의 온라인 소비행태는 크게 늘었다. 보험에서도 편리함, 신속함을 갖춘 디지털화 기대가 커지고 있다.

- 따라서, 신한금융그룹의 데이터를 활용하여, 새로운 개인 맞춤형 보험과 서비스를 제공하는 것을 목표로 하였다.

# 1.개요 (전체 구성 요약)

1) 신한라이프 데이터를 전처리 해서, 주력으로 개인화 보험 서비스를 만들 연금/저축 상품 결정

2) 신한은행과 신한증권에서 보험 서비스에 가입할 확률이 높은 연령대로 20-30대 선정

3) 2-30대가 주로 쓰는 카드 소비 형태 파악 후, 그 내용을 신규 보험 상품 개발에 반영

4) 20-30대가 보험사에 신규 계약 고객이 될 동향 데이터 모델링

5) 기존 보험 상품과 2-30대의 소비 스타일을 반영한 신규 보험 서비스 소개

# 2.데이터 전처리

- 데이터를 불러와 모든 칼럼 값이 0이 되는 행이 없는 것을 확인 후, 신한라이프에 해당하는 칼럼 만 추출하였다.

```
life_raw = data_clean.iloc[:,373:]
zero_rows = life_raw[(life_raw == 0).all(axis=1)]
life_raw # 신한 라이프의 원 데이터
```

| | la01r | la02r | la03r | la04r | la05r | la06r | la07r | lb01r | lb02r | lb03r | ... | le12r | le13r | le14r | lf01r | lf02r | lf03r | lf04r | lf05r | lf06r | lf07r |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.20 | 0.50 | 0.20 | 0.02 | 0.02 | 0.14 | 0.03 | 0.38 | 0.35 | 0.12 | ... | 0.27 | 0.14 | 0.04 | 0.75 | 0.71 | 0.24 | 0.15 | 0.19 | 0.48 | 0.19 |
| **1** | 0.21 | 0.45 | 0.34 | 0.00 | 0.01 | 0.07 | 0.00 | 0.41 | 0.33 | 0.12 | ... | 0.28 | 0.13 | 0.04 | 0.73 | 0.69 | 0.22 | 0.16 | 0.21 | 0.48 | 0.16 |
| **2** | 0.13 | 0.52 | 0.30 | 0.00 | 0.00 | 0.09 | 0.00 | 0.38 | 0.36 | 0.19 | ... | 0.20 | 0.12 | 0.03 | 0.73 | 0.74 | 0.15 | 0.15 | 0.18 | 0.55 | 0.11 |
| **3** | 0.27 | 0.53 | 0.21 | 0.00 | 0.00 | 0.10 | 0.00 | 0.40 | 0.39 | 0.13 | ... | 0.28 | 0.11 | 0.03 | 0.72 | 0.70 | 0.22 | 0.15 | 0.21 | 0.47 | 0.16 |
| **4** | 0.25 | 0.50 | 0.33 | 0.01 | 0.01 | 0.05 | 0.02 | 0.42 | 0.32 | 0.13 | ... | 0.28 | 0.13 | 0.04 | 0.73 | 0.71 | 0.22 | 0.16 | 0.19 | 0.47 | 0.18 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **114598** | 0.00 | 0.33 | 0.33 | 0.00 | 0.00 | 0.00 | 0.33 | 0.23 | 0.41 | 0.29 | ... | 0.21 | 0.09 | 0.41 | 0.38 | 0.62 | 0.03 | 0.32 | 0.32 | 0.35 | 0.00 |
| **114599** | 0.00 | 0.50 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 | 0.25 | 0.30 | 0.25 | ... | 0.15 | 0.15 | 0.50 | 0.60 | 0.60 | 0.10 | 0.20 | 0.35 | 0.40 | 0.05 |
| **114600** | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.19 | 0.43 | 0.19 | ... | 0.33 | 0.22 | 0.26 | 0.33 | 0.52 | 0.07 | 0.37 | 0.37 | 0.24 | 0.02 |
| **114601** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.25 | 0.50 | 0.33 | ... | 0.42 | 0.17 | 0.25 | 0.58 | 0.58 | 0.00 | 0.33 | 0.17 | 0.50 | 0.00 |
| **114602** | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.50 | 0.36 | ... | 0.17 | 0.07 | 0.47 | 0.33 | 0.37 | 0.07 | 0.47 | 0.33 | 0.17 | 0.03 |

114600 rows × 60 columns

- 신한라이프에 해당하는 데이터를 추출한 후, 그 데이터 프레임에도 모든 칼럼 값이 0이 되는 행이 없는 것을 확인 하였다.

```
zero_rows = life_raw[(life_raw == 0).all(axis=1)]
zero_rows
```

| la01r | la02r | la03r | la04r | la05r | la06r | la07r | lb01r | lb02r | lb03r | ... | le12r | le13r | le14r | lf01r | lf02r | lf03r | lf04r | lf05r | lf06r | lf07r |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

0 rows × 60 columns

# 2.데이터 전처리

- 신한 라이프 상품군 7가지 (종신 보험, 건강 보험, 상해 보험, 어린이 보험, 연금/저축 보험, 변액보험, 기타) 중 어떤 상품군이 가장 보험계약고객 비율이 높고, 보험계약월납보험료가 높은지 등의 분포를 파악하고자 하였다.

- 따라서, 상품 군 7가지에 대한 데이터를 보유한 '최근1년보험계약고객', '보험계약월납보험료', '보험금지급경험고객비율', '보험금지급경험고객비율', '보험계약대출경험고객비율' 이라는 5개의 신한 라이프 중분류를 선택하여 전처리를 진행하였다.

- 예시로, '최근1년보험계약고객' 중분류에 해당하는 7가지의 상품군 칼럼을 불러오고, 모든 칼럼 값이 0인 행은 제외한 새로운 데이터 프레임을 만들었다.

```
life_recent_s = life_raw.iloc[:, :7]
life_recent_s_nozero = life_recent_s[(life_recent_s != 0).any(axis=1)]
life_recent_s_nozero
```

- 위와 같이 5개의 중분류에 대해 똑같은 작업을 수행하여 5개의 새로운 데이터 프레임을 만들었다.

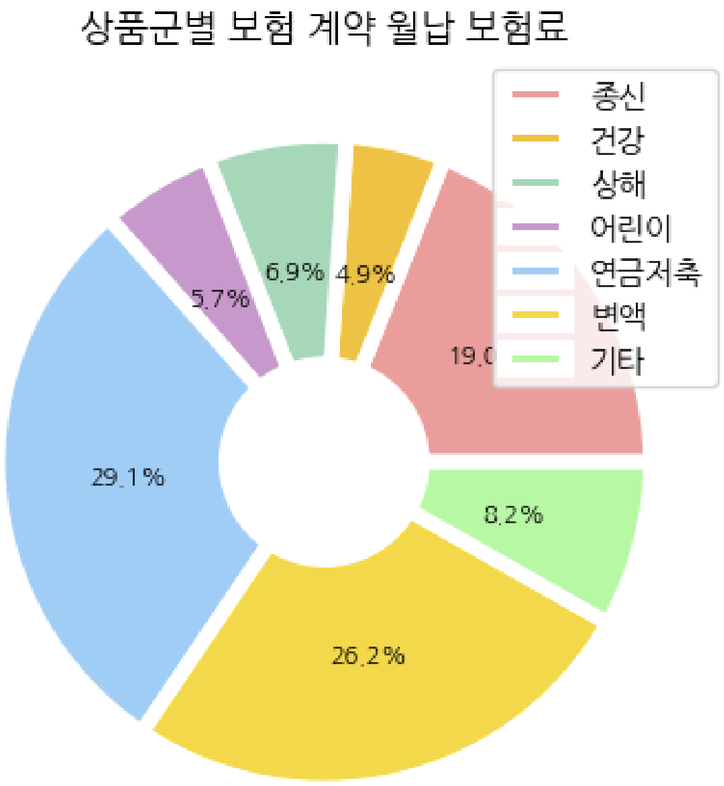
# 3.데이터 분석

5개의 새로운 데이터 프레임을 활용하여
다음과 같은 항목의 상품군별 비율을 시각화를 진행하였다.

- 최근 1년 신규 보험계약 고객 비율
- 보험계약 고객 비율
- 보험계약 평균 월납 보험료
- 보험금 지급경험 고객 비율
- 보험계약 대출경험 고객 비율

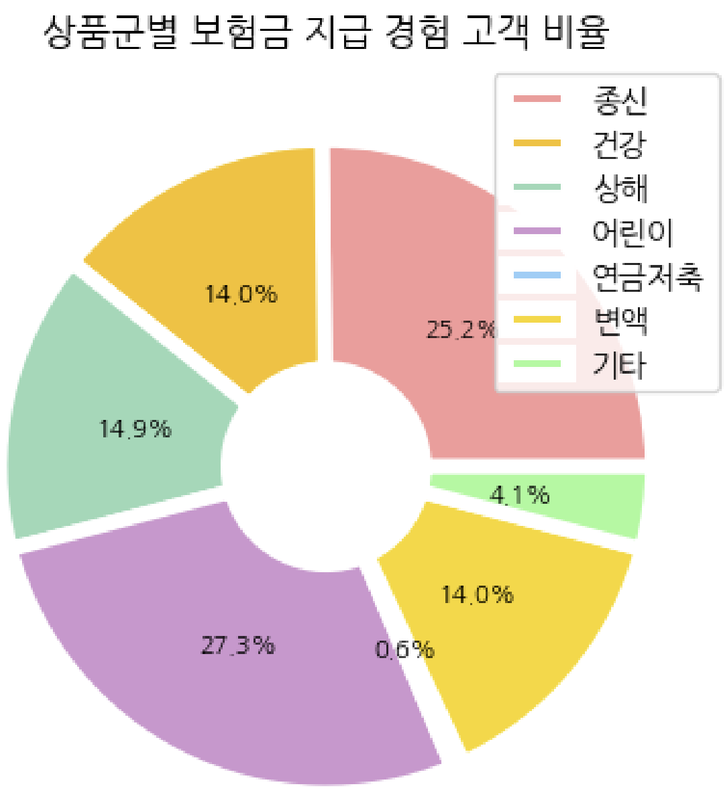
# 3.데이터 분석

## 보험계약 평균 월납 보험료




'연금저축' 상품이 월납 보험료
비율이 가장 높다.

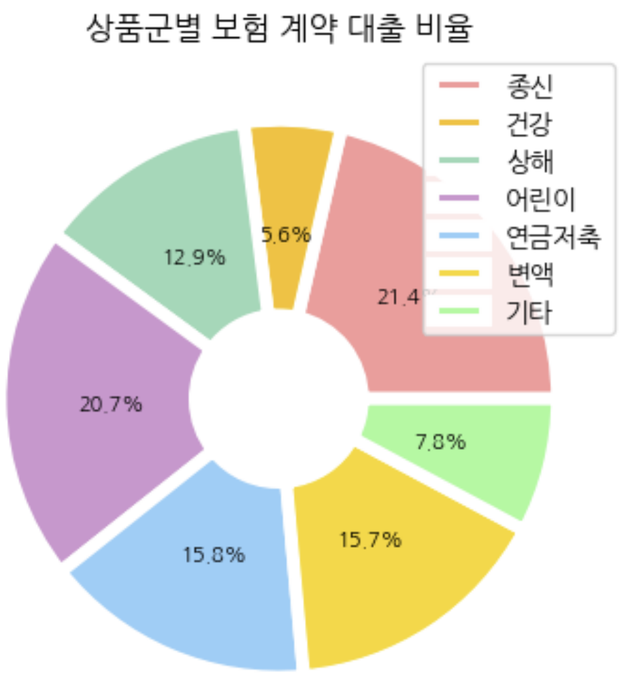
## 보험금 지급 경험 고객 비율




'연금저축' 보험금 지급
비율이 가장 낮다.

# 3.데이터 분석

## 보험 계약 대출

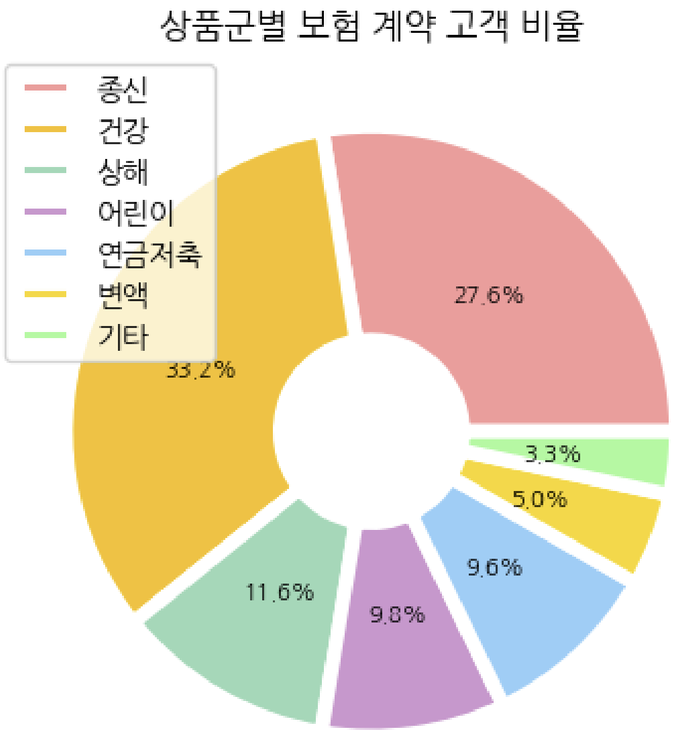



'연금저축' 상품이 보험 계약 대출 비율은 높은 편이다.

- 앞에 세 개의 차트 시각화를 통해 월납 보험료 비율이 가장 높고, 보험금 지급 비율이 가장 낮으며, 보험 계약 대출 비율은 높은 '연금/저축' 상품이 가격면에서 효율이 높다고 판단하였다.
- 하지만, 뒤에 두 시각화 차트를 통해 실제 가입 고객은 하위권이며 최근 1년 비율도 현저히 줄어드는 것을 발견 하였다.
- 따라서 연금/저축 보험을 주력 상품으로 설정하기로 결정하였다.

# 3.데이터 분석

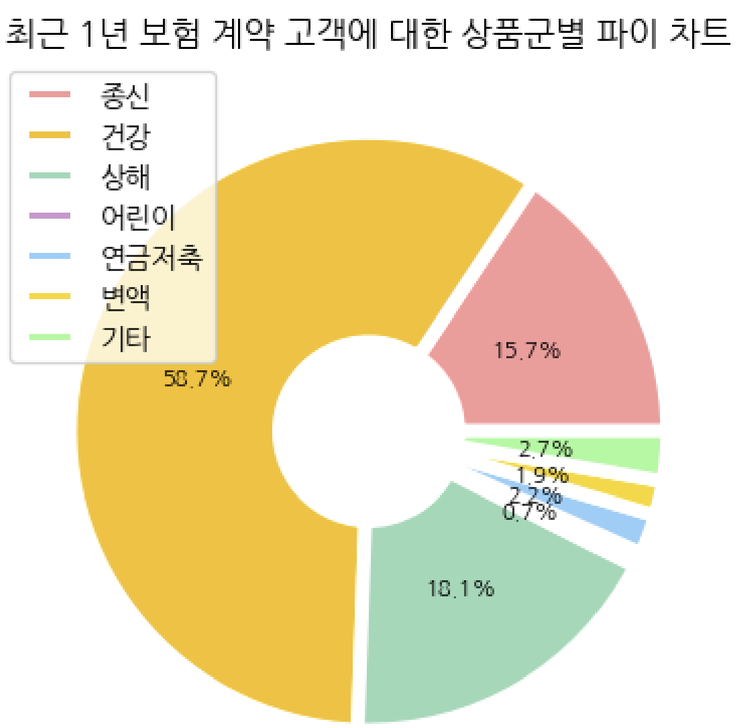
## 보험 계약 고객 비율




'연금저축' 상품이 보험 계약
고객 비율이 낮은 편이다.

## 최근 1년 신규 보험계약 고객 비율

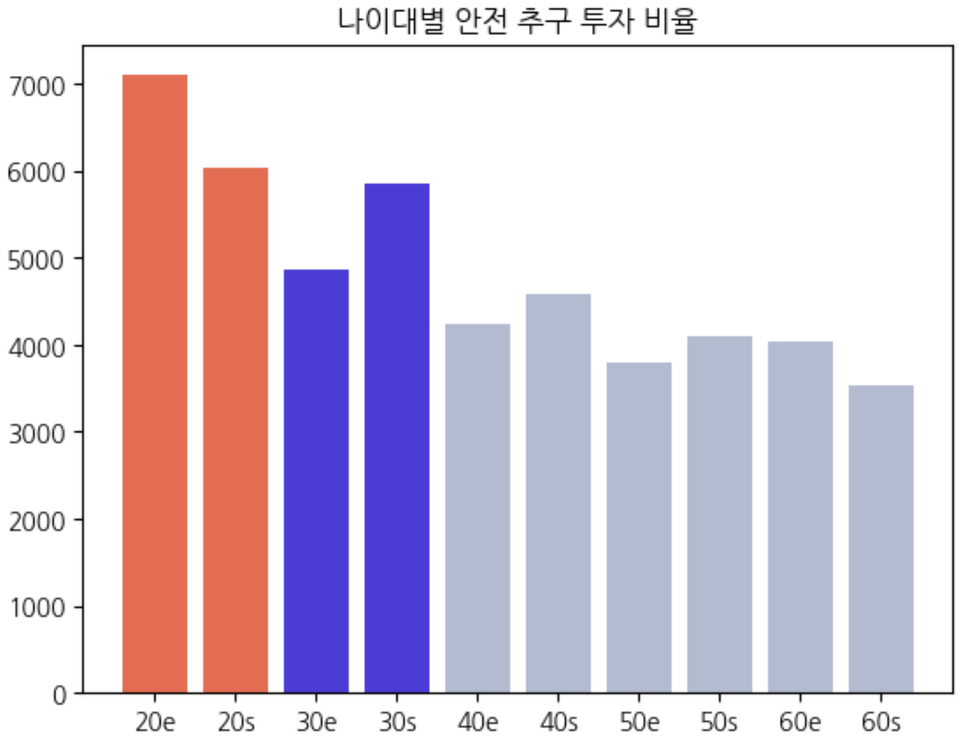
건강: 54614.88
상해: 16850.03
종신: 14572.09
기타: 2501.61
연금저축: 2002.36
변액: 1727.60
어린이: 695.43




'연금저축' 상품이 최근 1년
신규 보험 계약 고객 비율이 현저히 낮다.

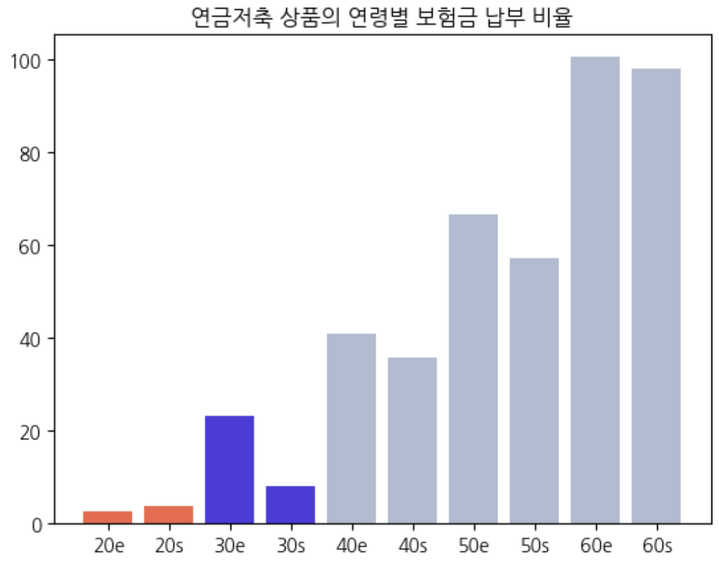
# 3.데이터 분석

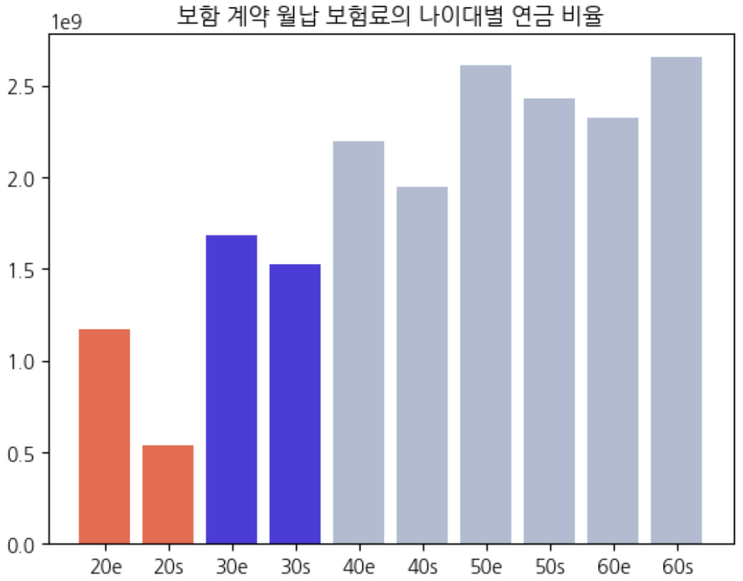
추가적으로, 신한은행과 신한투자증권 데이터 분석을 통해 연금/저축 상품을 들 가능성이 높은 나이대를 선별

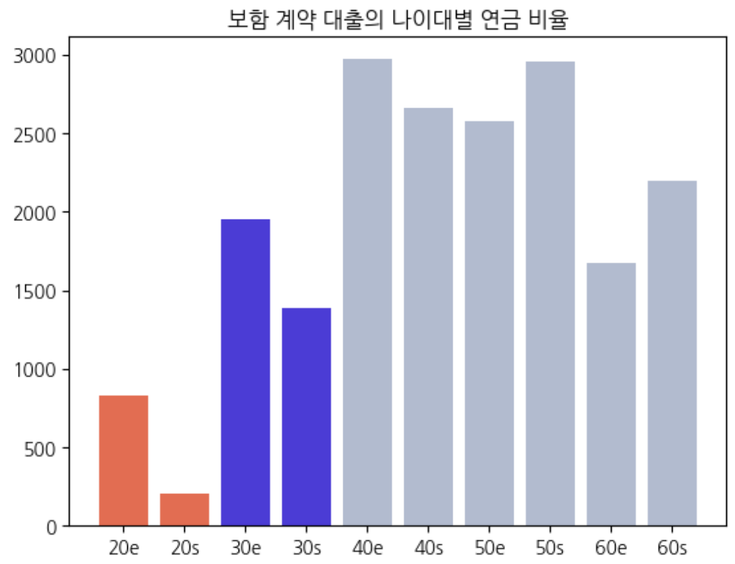



- '연금/저축' 상품에 들 법한 나이대를 추측하기 위해 나이대별 안전을 추구하는 비율을 산출하였다.
- 신한은행 데이터에서 유동성거래 중 '연금 관련 거래 비율' 컬럼과 신한투자증권의 투자 등급에서 '안전형'과 '안전추구형' 비율 컬럼을 추출하여 막대 그래프로 시각화하였다.
- 20-30대가 '연금/저축' 상품을 가입할 가능성이 높다고 판단하였다.

# 3.데이터 분석



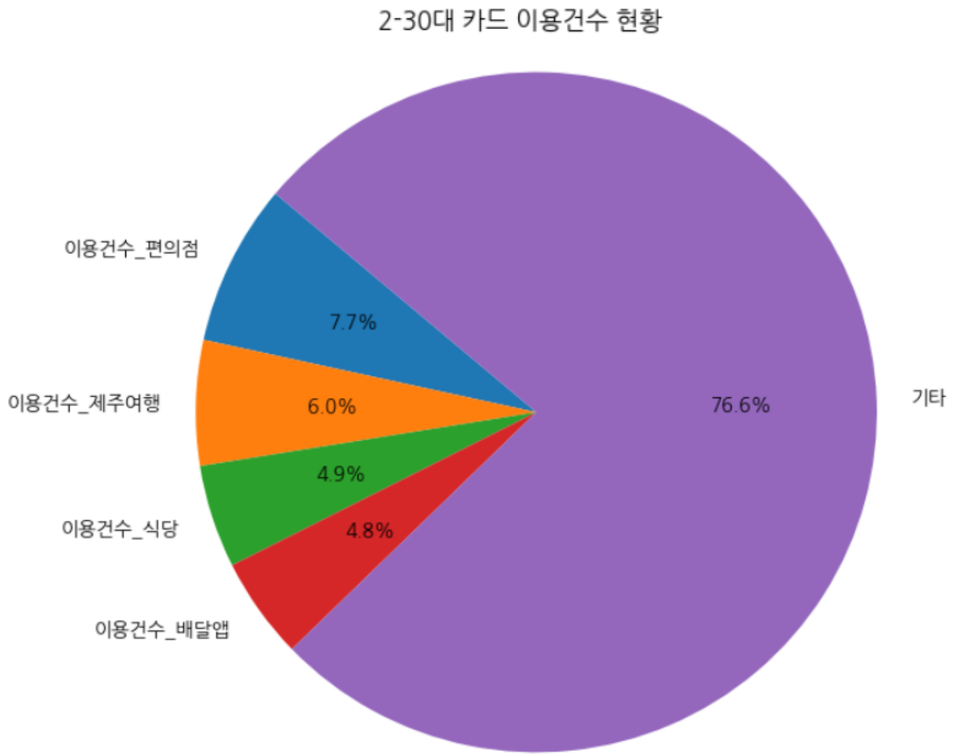

- '연금/저축' 상품의 나이대별 비율을 시각화 해본 결과 주력으로 하려는 20-30대 현황이 낮다는 것을 인지하였다.
- 상품을 개발할 때, 보험금 납부 비율과 보험료 납부 비율을 높일 수 있는 방안 모색 필요성을 인지하였다.
- 연금/저축 상품의 신규 보험계약 고객 비율을 높이기 위해서,20-30대를 연금/저축 보험 상품의 새로운 타켓 연령층으로 설정하였다.

# 3.데이터 분석

- 20-30대의 카드 이용건수가 가장 높은 4개의 항목에서 신한 카드 이용 실적이 일정 기준을 넘어서면 자동으로 연금/저축에 쓸 수 있는 포인트를 적립하는 서비스를 개발하고자 하였다.
- 따라서, 신한 카드사 데이터를 활용하여 20-30대의 카드 이용건수가 가장 높은 4개의 항목 1위) 편의점, 2위) 제주여행, 3위) 식당, 4위) 배달앱 이라는 것을 파악하였다.

```
# 가장 큰 합계를 가지는 칼럼 이름 출력 및 순위 표시
for rank, idx in enumerate(max_columns, 1):
    column_idx = 4 + idx
    column_name = selected_rows.columns[column_idx]
    original_column_name = column_mapping.get(column_name, column_name)
    print(f"{rank}위 가장 큰 합계를 가지는 칼럼:", original_column_name)
```

```
1위 가장 큰 합계를 가지는 칼럼: 이용건수_편의점
2위 가장 큰 합계를 가지는 칼럼: 이용건수_제주여행
3위 가장 큰 합계를 가지는 칼럼: 이용건수_식당
4위 가장 큰 합계를 가지는 칼럼: 이용건수_배달앱
```



# 4.데이터 모델링 및 예측

| | pk1 | pk3 | la05r |
|---|---|---|---|
| 0 | 2022-08-01 | 20e | 4.13 |
| 1 | 2022-08-01 | 20s | 4.08 |
| 2 | 2022-08-01 | 30e | 6.19 |
| 3 | 2022-08-01 | 30s | 4.43 |
| 4 | 2022-10-01 | 20e | 6.57 |
| 5 | 2022-10-01 | 20s | 3.11 |
| 6 | 2022-10-01 | 30e | 4.59 |
| 7 | 2022-10-01 | 30s | 5.03 |
| 8 | 2022-12-01 | 20e | 8.63 |
| 9 | 2022-12-01 | 20s | 5.79 |
| 10 | 2022-12-01 | 30e | 7.90 |
| 11 | 2022-12-01 | 30s | 9.61 |
| 12 | 2023-02-01 | 20e | 10.30 |
| 13 | 2023-02-01 | 20s | 7.61 |
| 14 | 2023-02-01 | 30e | 11.92 |
| 15 | 2023-02-01 | 30s | 9.20 |
| 16 | 2023-04-01 | 20e | 13.35 |
| 17 | 2023-04-01 | 20s | 9.11 |
| 18 | 2023-04-01 | 30e | 14.33 |
| 19 | 2023-04-01 | 30s | 10.75 |
| 20 | 2023-06-01 | 20e | 14.84 |
| 21 | 2023-06-01 | 20s | 10.61 |
| 22 | 2023-06-01 | 30e | 13.37 |
| 23 | 2023-06-01 | 30s | 9.16 |

<merge_data_new>

- 향후 1년의 연금/저축 상품에 대한 계약 고객 비율 수치 예측하기 위해 왼쪽 그림과 같이 데이터 전처리를 하였다.

- 날짜 : 기존의 표시된 날짜 방식에서 datetime형으로 수정하였다.
- 연령 : 필요한 연령만 색출하였다. (20-30대)
- 수치 : 해당 날짜와 연령에 대한 비율 수치 값들의 합으로 구성하였다.
- 변수 : pk1 : 년-월 날짜 데이터, pk3 : 연령 데이터, la05r : 최근 1년 연금/저축 상품 계약 고객 비율

# 4.데이터 모델링 및 예측

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from datetime import datetime

data = merged_data_new

data['year'] = pd.to_datetime(data['pk1']).dt.year
data['month'] = pd.to_datetime(data['pk1']).dt.month

data_enconded = pd.get_dummies(data, columns=['pk3'], prefix=['age'])

features = ['year', 'month', 'age_20e', 'age_20s', 'age_30e', 'age_30s']
target = 'la05r'

X = data_enconded[features]
y= data_enconded[target]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

predicted_values = []
prediction_dates = [datetime(2023, 8, 1), datetime(2023, 10, 1),datetime(2023, 12, 1), datetime(2024, 2, 1), datetime(2024, 4, 1),datetime(2024, 6, 1)]
for prediction_date in prediction_dates :
    future_year = prediction_date.year
    future_month = prediction_date.month

    future_data = pd.DataFrame([[future_year, future_month, 1, 0, 0, 0]], columns=features) # 1은 20대를 나타냄 / 2023년 8월 20대
    predicted_value = model.predict(future_data)
    predicted_values.append(predicted_value)
    print(f"{future_year}년 {future_month}월 예측 수치값: ", predicted_value)
```

```
2023년 8월 예측 수치값:  [9.4608]
2023년 10월 예측 수치값:  [9.2215]
2023년 12월 예측 수치값:  [9.761]
2024년 2월 예측 수치값:  [10.8895]
2024년 4월 예측 수치값:  [11.245]
2024년 6월 예측 수치값:  [13.0189]
```

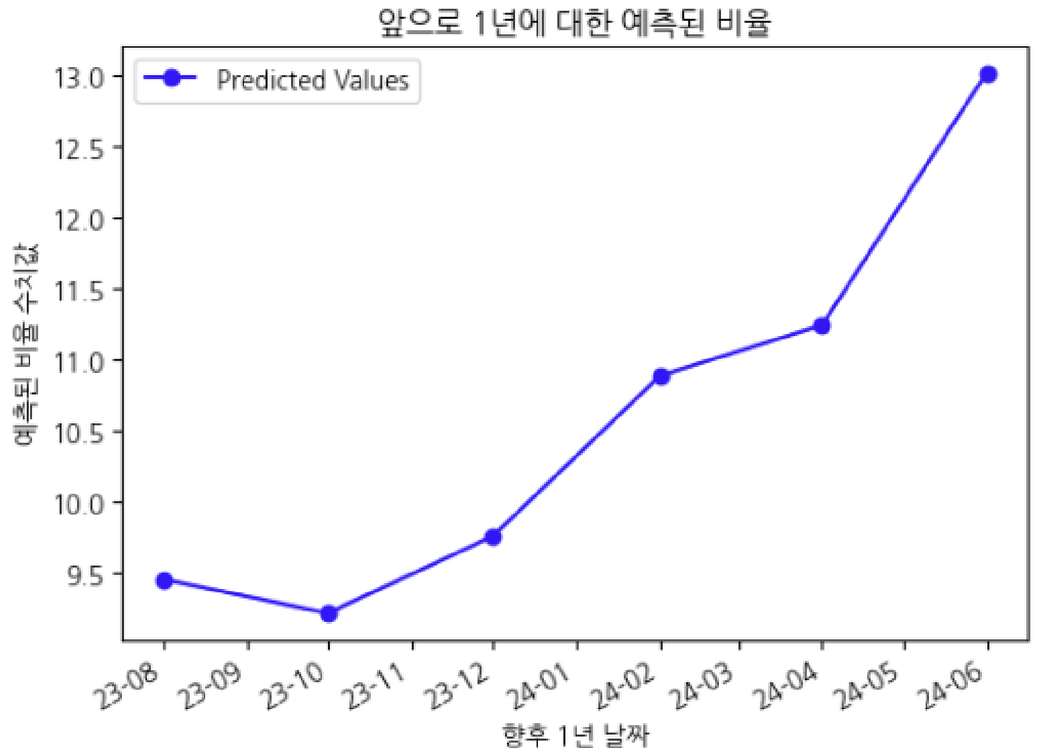
- 앞에 전처리된 데이터로 모델링하여 향후 1년의 연금/저축 상품에 대한 계약 고객 비율 수치 예측하였다.
- 회귀에 RandomForestRegressor 모델을 사용하였다.

-> 일반화 및 성능이 우수하며 파라미터 조정이 용이하다.

-> 사용한 데이터는 차원이 크지 않고 밀집된 데이터를 사용하므로 채택하였다.

# 4.데이터 모델링 및 예측

```
import matplotlib.pyplot as plt
import matplotlib.dates as mdates

dates = ['2023-08', '2023-10', '2023-12', '2024-02', '2024-04', '2024-06']

y = [predicted_values]

dates = [datetime.strptime(date, "%Y-%m") for date in dates]

plt.plot(dates, predicted_values, marker='o', color='b', label='Predicted Values')

plt.gca().xaxis.set_major_formatter(mdates.DateFormatter('%y-%m'))
plt.gca().xaxis.set_major_locator(mdates.MonthLocator())
plt.gcf().autofmt_xdate()

plt.title('앞으로 1년에 대한 예측된 비율')
plt.xlabel('향후 1년 날짜')
plt.ylabel('예측된 비율 수치값')
plt.legend()
plt.show()
```




- 최근 1년 계약 고객 비율을 통해 향후 1년을 예측한 값의 경향성을 보기 위해서 matplotlib.pyplot으로 꺾은 선 그래프로 시각화 하였다.
- 20-30대 '연금/저축' 상품 가입 비율이 올라갈 거라고 예측하였다.

# 4.데이터 모델링 및 예측

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error

data = merged_data_new

data['year'] = pd.to_datetime(data['pk1']).dt.year
data['month'] = pd.to_datetime(data['pk1']).dt.month

data_enconded = pd.get_dummies(data, columns=['pk3'], prefix=['age'])

features = ['year', 'month', 'age_20e', 'age_20s', 'age_30e', 'age_30s']
target = '1a05r'

X = data_enconded[features]
y= data_enconded[target]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)

predictions = model.predict(X_test)

mse = mean_squared_error(y_test, predictions)
print("Mean Squared Error : ", mse)
```

```
Mean Squared Error :  5.1913140738105215
```

- 예측 성능 평가에 MSE(평균제곱오차) 사용

-> 이상치에 취약하지만, 전처리를 통해 이상치를 제거하였으므로 채택

-> 예측 경향성이 아닌 얼마나 정확한지만 확인할 예정이므로 채택

- 결과 값은 5점 초반대로 낮아서 예측에 신빙성을 더하였다.

# 5.신사업 개요

## 주 이용고객은 20대-30대

앞에 신한 라이프 데이터 시각화와 신한 증권 및 은행 데이터 분석을 통해서 '연금/저축' 상품을 중점으로 하는 신사업의 주요 타깃은 20-30대로 확정지었다.

## 주 구매 동향 요식업 및 여행

'타깃층의  주 소비 품목으로 편의점, 제주여행, 배달앱, 식당 등 요식업 관련 소비 품목이 다수였다. 이에 요식업 관련 혜택을 제공하는 결합 상품을 신사업 아이템으로 채택하였다.

# 5.신사업 [청년 연금저축보험]


## 3新 지키미

건강 지키고
돈 지키고
혜택 지키자!

## 주 대상 2-30대

신한카드 전월 평균실적 이상&
편의점, 배달앱, 식당 결제 시

**자동 연금저축보험**
**적립금 캐시백 지급!**

# 5.신사업 [청년 연금저축보험]



# 5.신사업 [청년 연금저축보험]


## SNS Event1.
## 인스타그램 스토리
## @태그 이벤트

[신한 라이프] 청년연금저축 적립금
화면 캡쳐 후, 스토리 공유 시
랜덤으로 **더블 적립금 캐시백 지급!**

## SNS Event2.
## 친구추천 이벤트

청년연금저축 보험 가입시
기 가입자 친구 이름 입력하면
**추가 적립금 캐시백 지급!**

# 5.[청년 연금저축보험] 기대효과

## 자동 연금저축 캐시백 지급

신한카드 전월 실적 평균 이상&
편의점, 배달앱, 식당 결제시

카드사와는 별도로
자동 연금저축보험 캐시백 지급!

➜ 20-30대 고객 유입 증가

## SNS 활용 홍보기회 증대

분기별 SNS 이벤트 시행
인스타그램 이벤트
친구추가 이벤트

➜ 20-30대 고객 홍보기회 증대

## 청년 연금저축 보험

사회초년생 및
안정형 고객 대상
금융상품 제공
&
신한라이프 유입 고객 확대

감사합니다

SHINHAN LIFE
BLUE LIFE