통신사 고객 이탈 예측

기계학습 Team Project Proposal

팀명 : 이건희

2019204045 윤서환
2019204094 이건희
2019204030 이승훈
2019204083 김효준
2019204023 윤성호

Contents

주제 선정 배경 및 중요성

씨티 "전기·가스요금 인상에도 5월 물가상승률 3~3.5% 전망"

4인 가구 5G 요금 30만원… “통신비만 잡아도 물가 낮춘다”

4인 가구 기준, 5G 110GB 요금제 살펴보면
SKT·KT, 27만6000원…LGU+, 30만원
4인 가구 월 전기료 5만원보다 6배 비싸
요금제 구성도 10GB·110GB 두개뿐

이통 이용자 절반, 가입 통신사에 불만···5G 만족도 LTE보다 낮아

송고시간 | 2022-09-12 07:00

요약

조승한 기자
기자 페이지

| 알뜰폰 이용자 만족도 63%…이통사 이용자보다 높아

정부 “3월 물가 더 오를수도…알뜰폰 요금 추가 인하”

주제 선정 배경 및 중요성

출처: 한국 신용 평가

3) MVNO[1](이하 '알뜰폰') 점유율 상승이 통신사에 미치는 영향과 모니터링 요소는?

당분간 알뜰폰 점유율은 상승흐름을 보일 전망이나, 중간요금제 출시 등을 통한 통신사의 중·저가 요금 수요층 흡수, 알뜰폰 사업자의 제한된 경쟁력 등을 감안할 때, 알뜰폰 점유율 상승에도 통신 3사의 시장지위 및 이익창출력이 크게 훼손되지는 않을 것으로 본다.

다만, 정책지원 등으로 MVNO 사업자의 5G 요금경쟁력이 제고될 경우, 알뜰폰 시장 잠식에 따른 영향은 다소 커질 수 있다.

22년 고객용 휴대폰 시장 내 알뜰폰 점유율 2.1%p 상승

알뜰폰 가입자 수는 2021년 말 최초로 1천만명에 도달하였고, 2022년 말에는 약 1.3천만명까지 증가하였다. 이에 2019년까지 10% 내외에 머물던 점유율도 2022년 말 16.9%까지 상승하였다.

다만, 2019년 이후 알뜰폰 가입자 증가는 단말장치(태블릿PC, 웨어러블 기기 등)와 사물지능통신(차량관제, 원격관제 등) 회선증가에 기인하고 있으며, ARPU가 현저히 높은 고객용 휴대폰 시장 점유율은 통신 3사의 저가 요금제 출시 등으로 2021년까지 오히려 소폭 저하된 모습이었다.

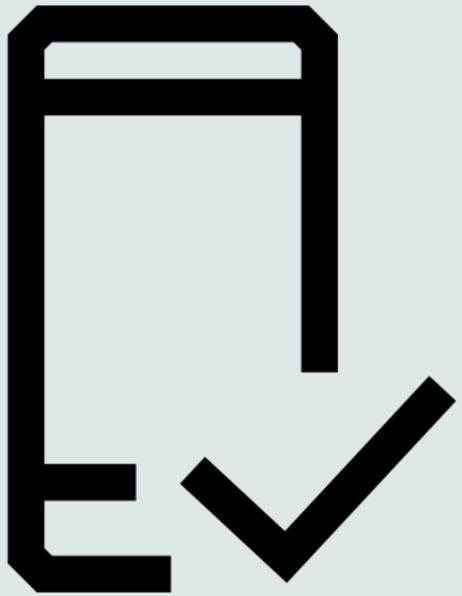
2022년 들어서는 고객용 휴대폰 시장에서도 알뜰폰 점유율이 상승 전환하였다(고객용 휴대폰 기준 MVNO 가입자 수 2021년 말 6.1백만명 → 2022년 말 7.3백만명). 이는 5G 출시로 통신 3사 주력 요금제가 비싸진 가운데, KB국민은행 등 자금력과 서비스 역량을 갖춘 사업자들이 시장에 진입하면서 가입자 유입에 영향을 준 것으로 보인다.

주제 선정 배경 및 중요성

이탈 고객 행동 및 동인 이해와 식별 필요

이탈 고객 사전 예방 필요

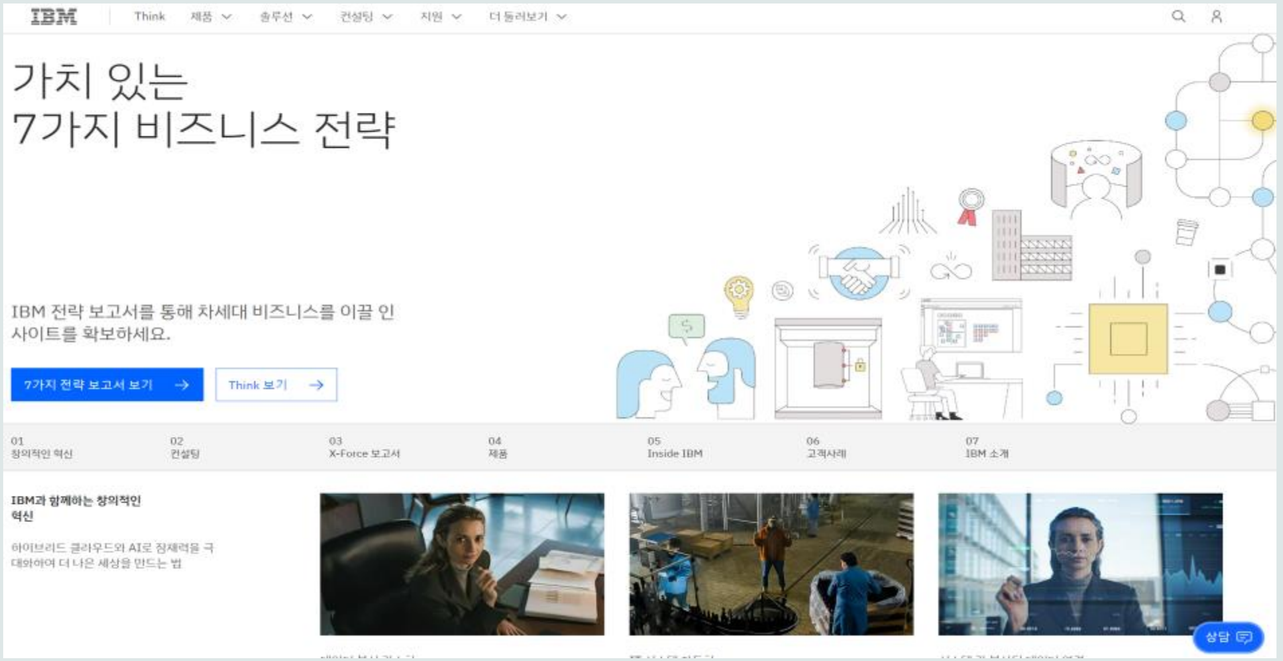
통신사 고객 유지 및 서비스 향상에 기여



데이터 설명

https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113

IBM의 캘리포니아 거주 고객 약 7043명의 전화 및 인터넷 서비스 데이터



데이터 설명 - 컬럼

- CustomerID: 고유 ID
- City: 도시
- Zip Code: 우편번호
- Latitude: 위도
- Longitude: 경도
- Gender: 성별
- Senior Citizen: 65세 이상인지 표시
- Partner: 커플 여부
- Dependents: 가족과 함께 살고 있는지 여부
- Tenure Months: 통신사 가입 기간
- Phone Service: 전화 서비스 가입 여부
- Multiple Lines: 여러 전화 회선에 가입했는지 여부
- Internet Service: 회사에서 인터넷 서비스에 가입했는지 여부
- Online Security: 온라인 보안 서비스에 가입했는지 여부

데이터 설명 - 컬럼

- Online Backup: 온라인 백업 서비스에 가입된 여부
- Device Protection: 기기 보호 요금제 가입 여부- Tech Support: 회사의 추가 기술 지원 계획에 가입 여부
- Streaming TV: TV 서비스 여부
- Streaming Movies: 영화를 구독 서비스 여부
- Contract: 계약 기간
- Paperless Billing: 인터넷 영수증 발급 여부
- Payment Method: 결제하는 방법
- Monthly Charge: 월 통신 이용 비용
- Total Charges: 분기 말까지 계산된 고객 총 요금
- Churn Label: 이번 분기에 회사를 떠났는지 여부(Yes, No)
- Churn Value: 이번 분기에 회사를 떠났는지 여부(1: Yes, 2: No)
- Churn Score: 이탈할 가능성을 나타내는 점수
- CLTV: 고객가치
- Churn Reason: 이탈하는 사유

데이터 설명 - 컬럼

전체 데이터 type과 cardinality 확인

	Column	d_type	unique_sample	n_uniques
0	CustomerID	object	[3668-QPYBK, 9237-HQITU, 9305-CDSKC, 7892-POOK...	7043
1	Count	int64	[1]	1
2	Country	object	[United States]	1
3	State	object	[California]	1
4	City	object	[Los Angeles, Beverly Hills, Huntington Park, ...	1129
5	Zip Code	int64	[90003, 90005, 90006, 90010, 90015]	1652
6	Lat Long	object	[33.964131, -118.272783, 34.059281, -118.30742...	1652
7	Latitude	float64	[33.964131, 34.059281, 34.048013, 34.062125, 3...	1652
8	Longitude	float64	[-118.272783, -118.30742, -118.293953, -118.31...	1651
9	Gender	object	[Male, Female]	2
10	Senior Citizen	object	[No, Yes]	2
11	Partner	object	[No, Yes]	2
12	Dependents	object	[No, Yes]	2
13	Tenure Months	int64	[2, 8, 28, 49, 10]	73
14	Phone Service	object	[Yes, No]	2
15	Multiple Lines	object	[No, Yes, No phone service]	3
16	Internet Service	object	[DSL, Fiber optic, No]	3
17	Online Security	object	[Yes, No, No internet service]	3
18	Online Backup	object	[Yes, No, No internet service]	3
19	Device Protection	object	[No, Yes, No internet service]	3
20	Tech Support	object	[No, Yes, No internet service]	3
21	Streaming TV	object	[No, Yes, No internet service]	3
22	Streaming Movies	object	[No, Yes, No internet service]	3
23	Contract	object	[Month-to-month, Two year, One year]	3
24	Paperless Billing	object	[Yes, No]	2
25	Payment Method	object	[Mailed check, Electronic check, Bank transfer...	4
26	Monthly Charges	float64	[53.85, 70.7, 99.65, 104.8, 103.7]	1585
27	Total Charges	object	[108.15, 151.65, 820.5, 3046.05, 5036.3]	6531
28	Churn Label	object	[Yes, No]	2
29	Churn Value	int64	[1, 0]	2
30	Churn Score	int64	[86, 67, 84, 89, 78]	85
31	CLTV	int64	[3239, 2701, 5372, 5003, 5340]	3438
32	Churn Reason	object	[Competitor made better offer, Moved, Competit...	20

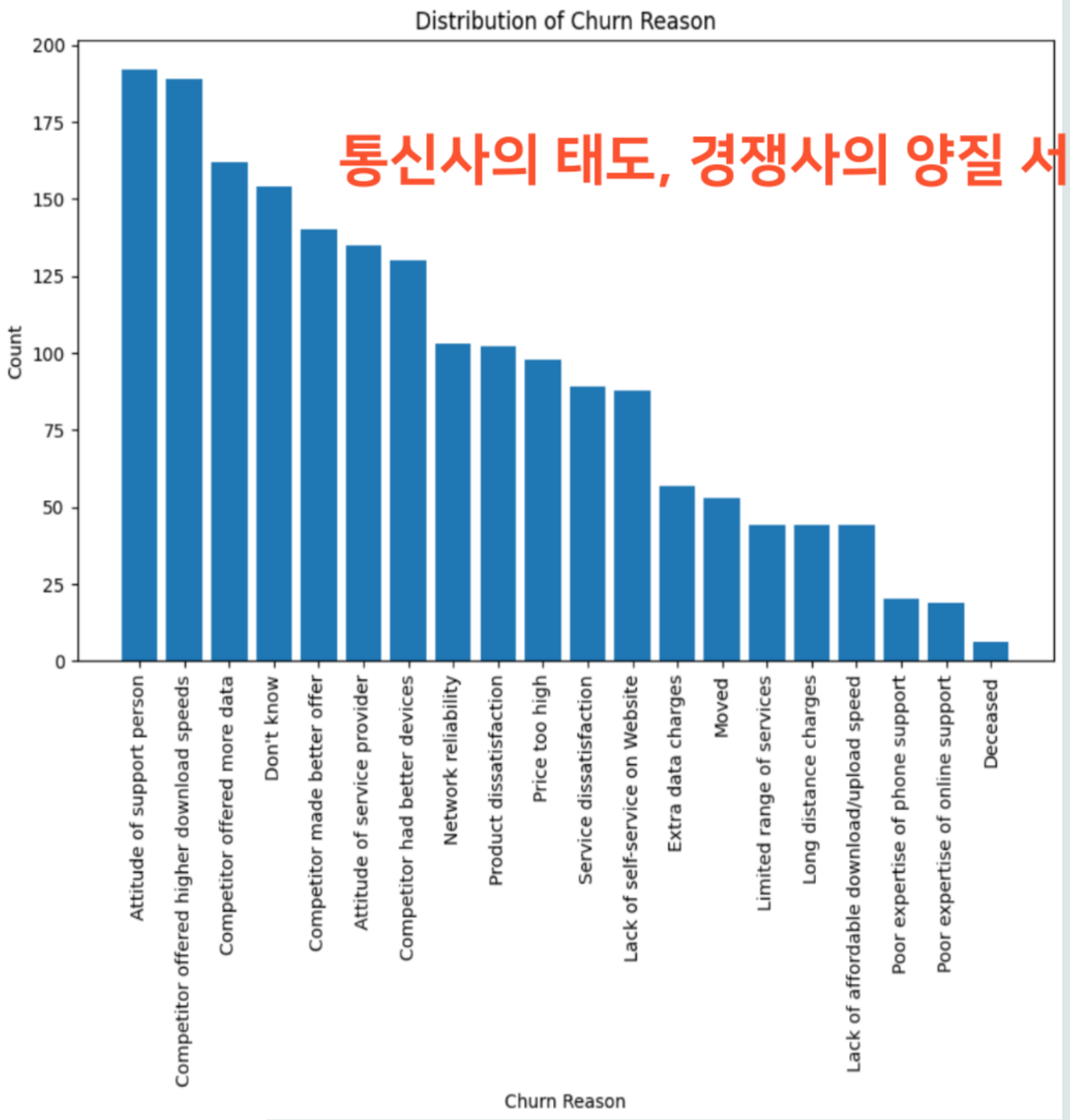
	Count	Zip Code	Latitude	Longitude	Tenure Months	Monthly Charges	Churn Value	Churn Score	CLTV
count	7043.0	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000
mean	1.0	93521.964646	36.282441	-119.798880	32.371149	64.761692	0.265370	58.699418	4400.295755
std	0.0	1865.794555	2.455723	2.157889	24.559481	30.090047	0.441561	21.525131	1183.057152
min	1.0	90001.000000	32.555828	-124.301372	0.000000	18.250000	0.000000	5.000000	2003.000000
25%	1.0	92102.000000	34.030915	-121.815412	9.000000	35.500000	0.000000	40.000000	3469.000000
50%	1.0	93552.000000	36.391777	-119.730885	29.000000	70.350000	0.000000	61.000000	4527.000000
75%	1.0	95351.000000	38.224869	-118.043237	55.000000	89.850000	1.000000	75.000000	5380.500000
max	1.0	96161.000000	41.962127	-114.192901	72.000000	118.750000	1.000000	100.000000	6500.000000

수치형 데이터 describe

데이터 설명

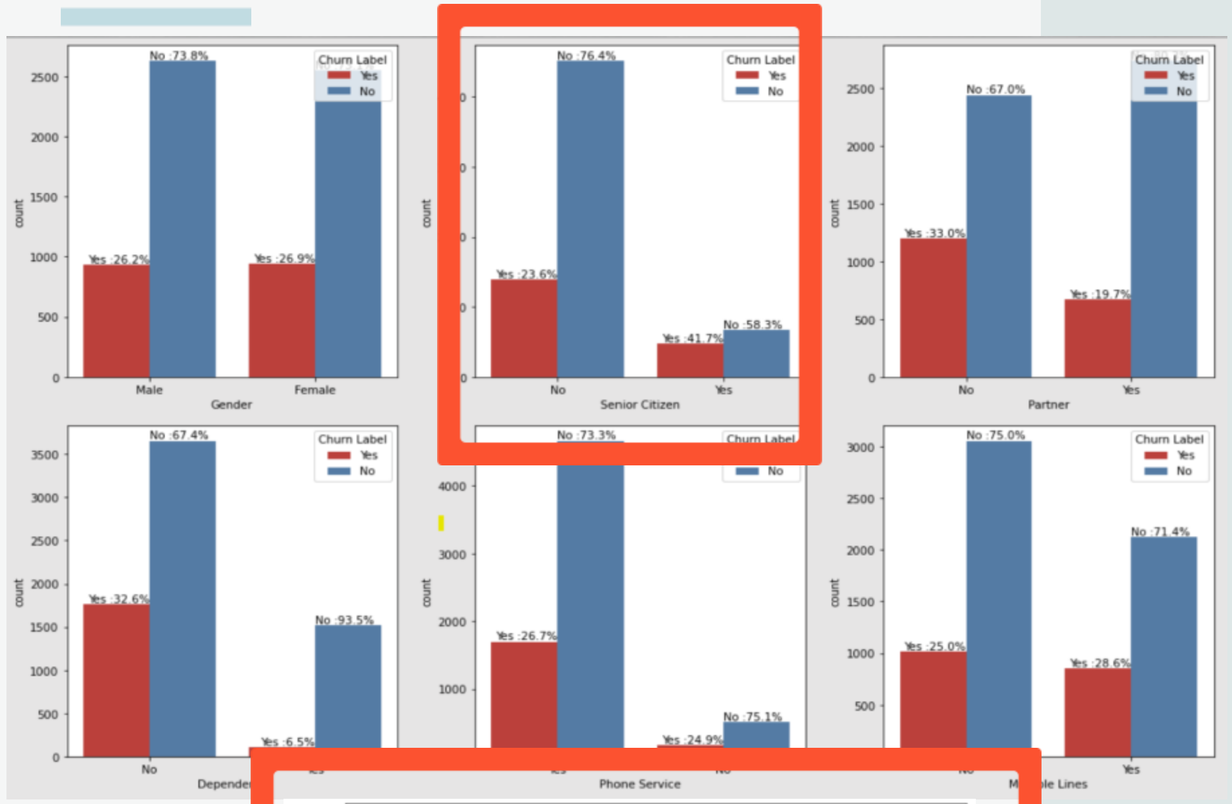
CustomerID	0
Count	0
Country	0
State	0
City	0
Zip Code	0
Lat Long	0
Latitude	0
Longitude	0
Gender	0
Senior Citizen	0
Partner	0
Dependents	0
Tenure Months	0
Phone Service	0
Multiple Lines	0
Internet Service	0
Online Security	0
Online Backup	0
Device Protection	0
Tech Support	0
Streaming TV	0
Streaming Movies	0
Contract	0
Paperless Billing	0
Payment Method	0
Monthly Charges	0
Total Charges	0
Churn Label	0
Churn Value	0
Churn Score	0
CLTV	0
Churn Reason	5174

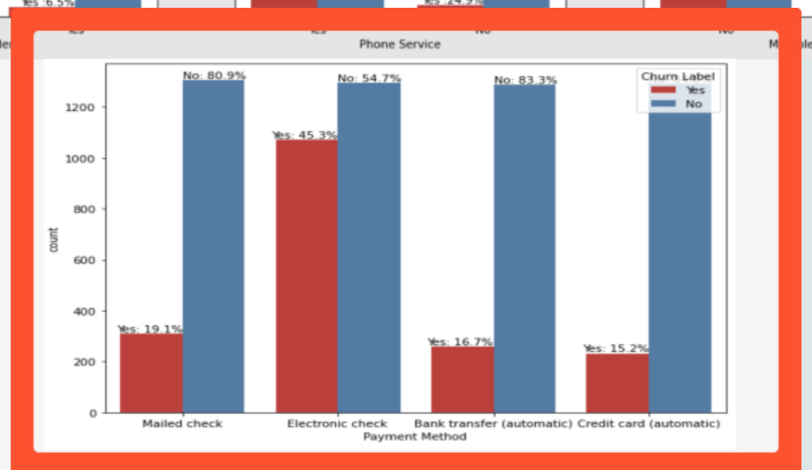
dtype: int64

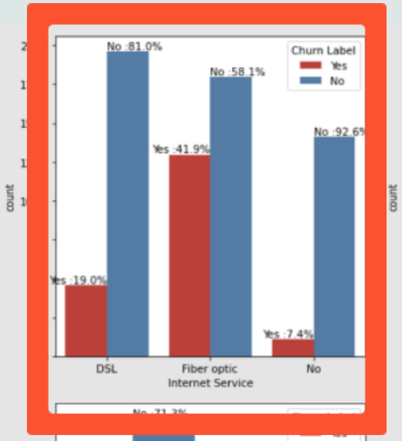


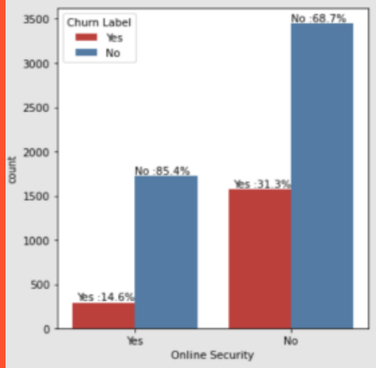
통신사의 태도, 경쟁사의 양질 서비스가 주된 이탈 사유

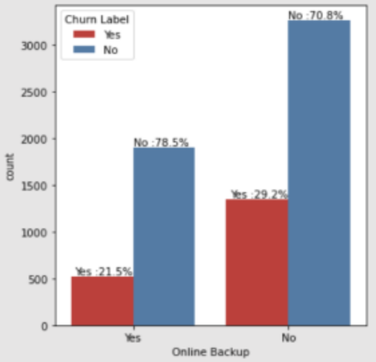
데이터 EDA - 변수에 따른 이탈률(범주형)

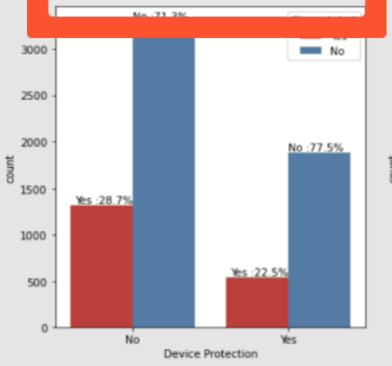


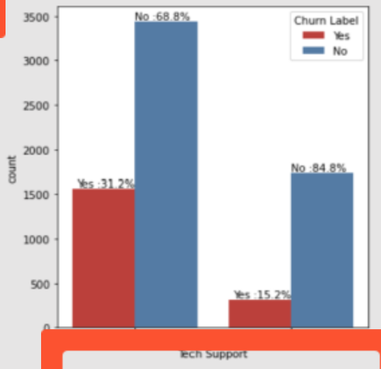


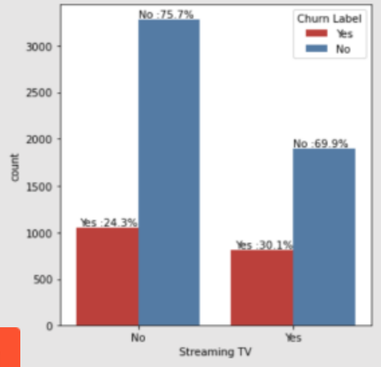


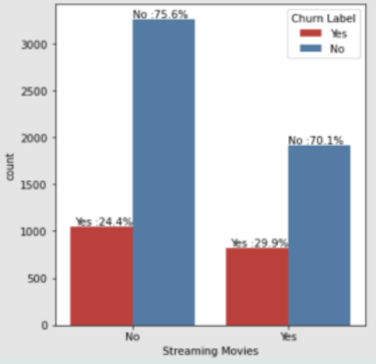


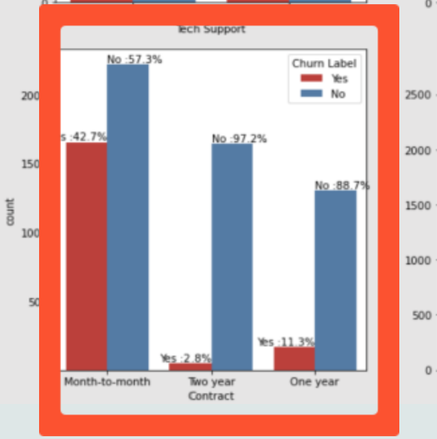


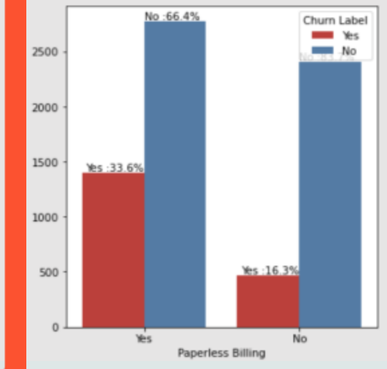






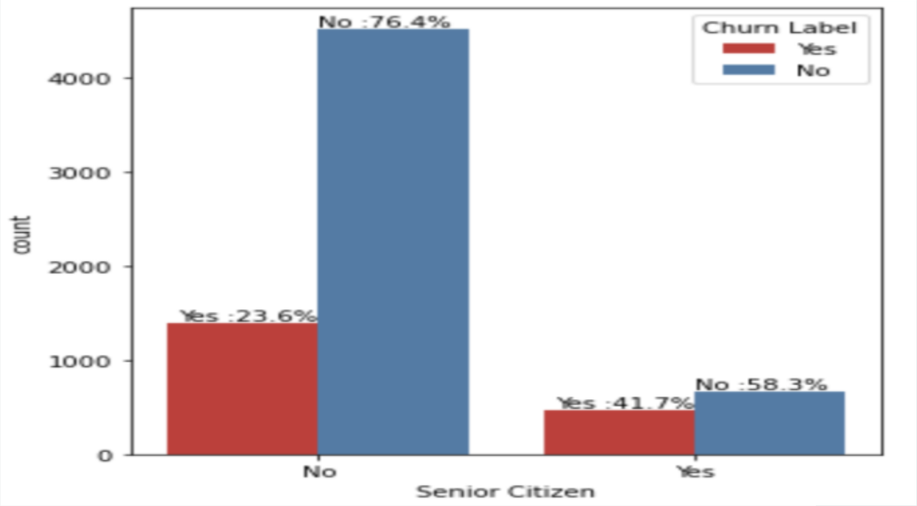


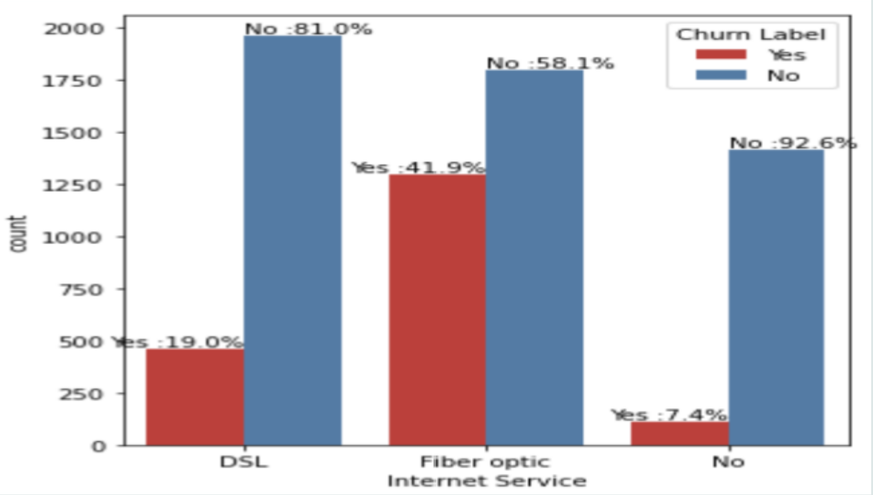


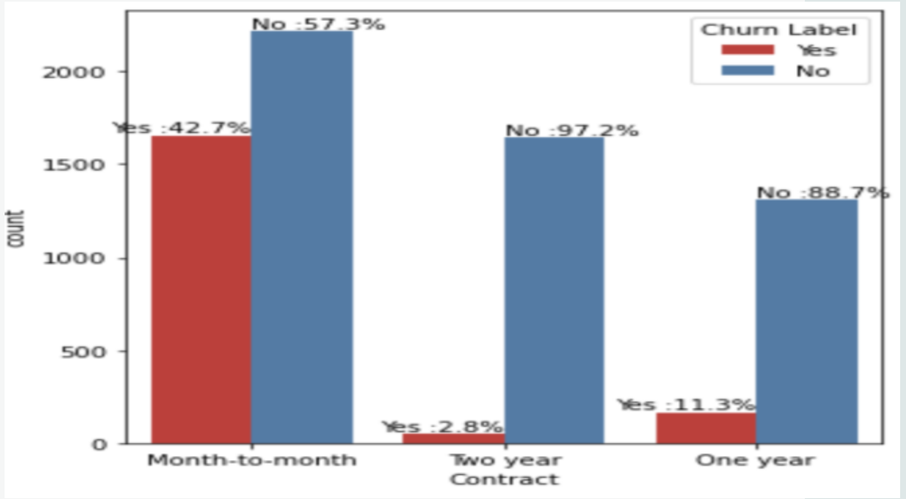


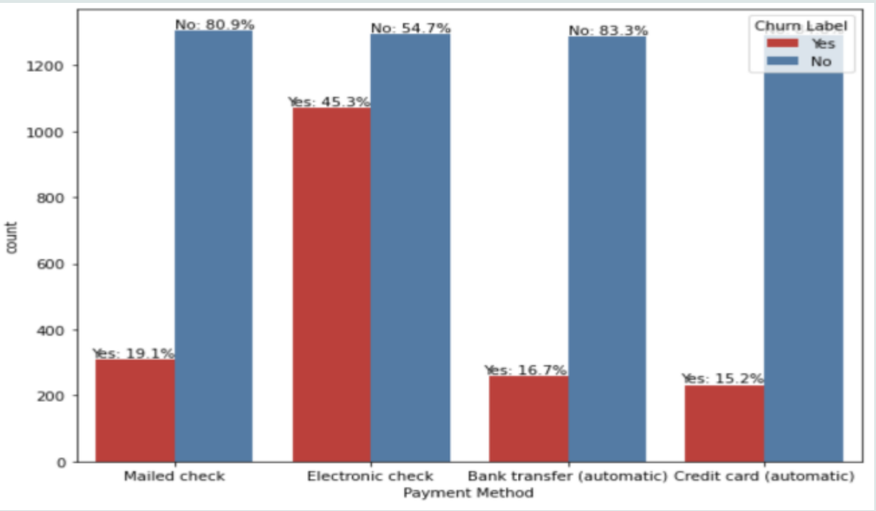
데이터 EDA - 변수에 따른 이탈률(범주형)

영향력이 클 것으로 예상되는 변수 4개 확인



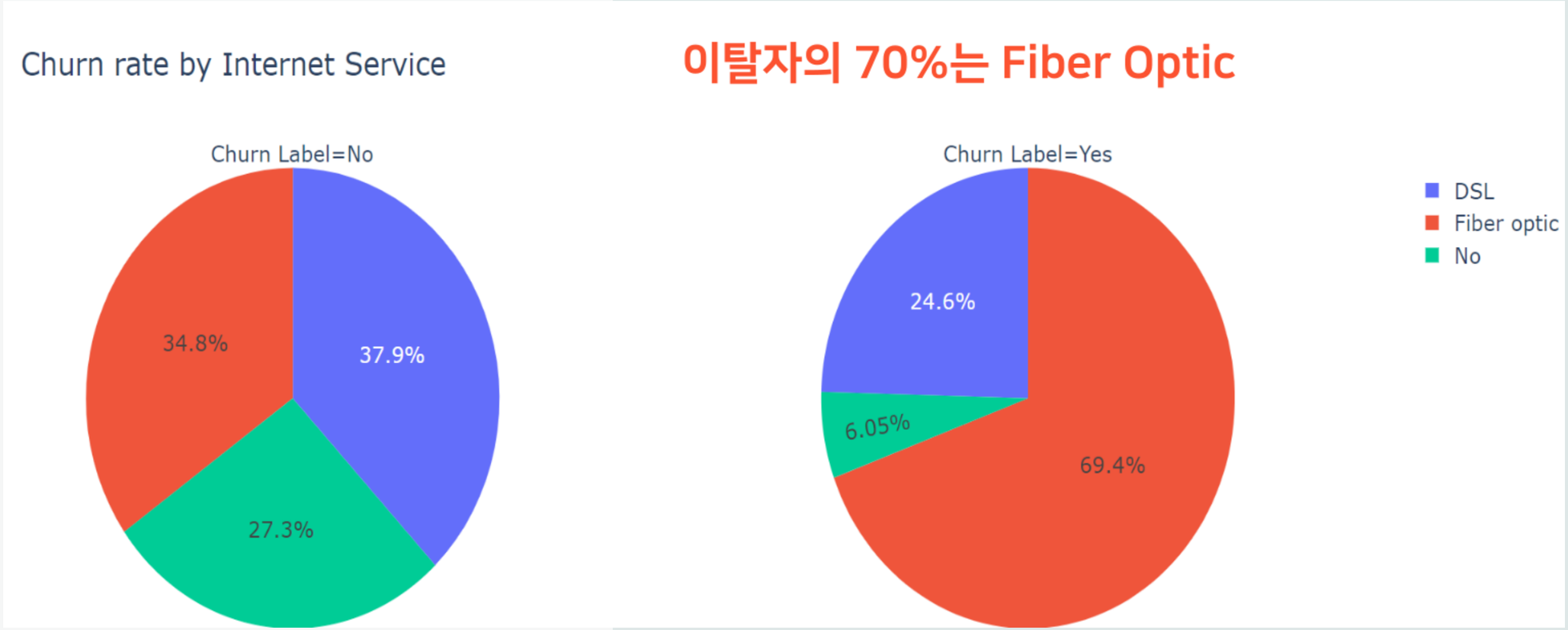






데이터 EDA - 주요 변수(범주형)

1. Internet Service

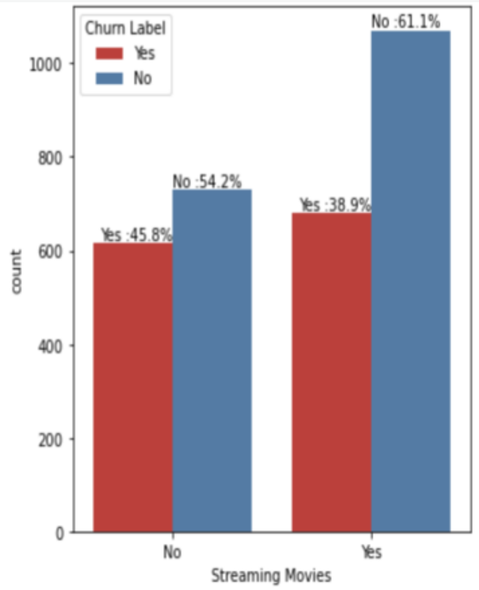


데이터 EDA - 주요 변수(범주형)

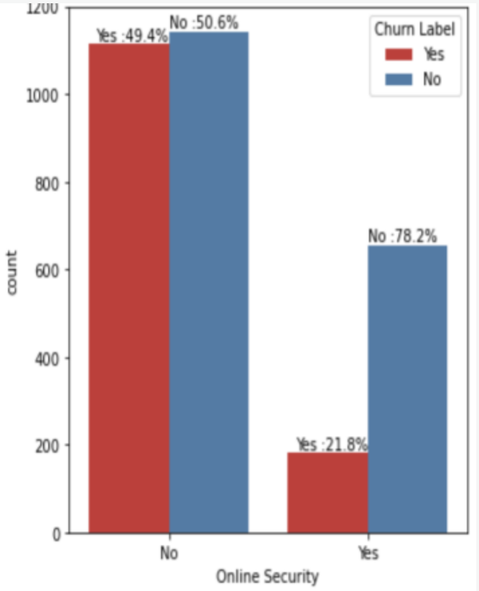
1. Internet Service(Fiber Optic)

— Yes
— No

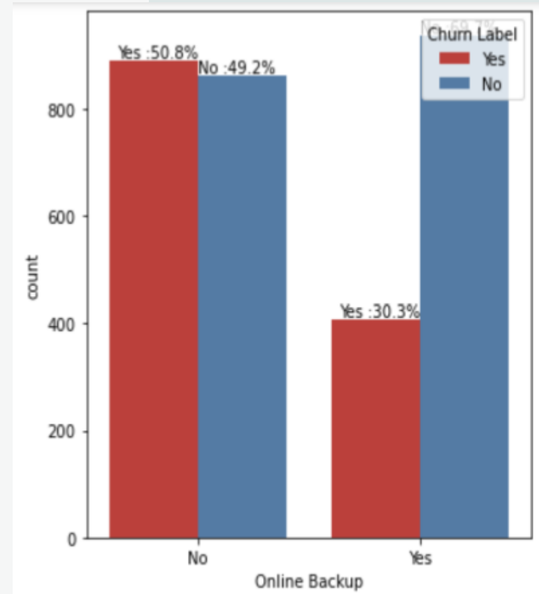
인터넷 서비스 연관 변수 이탈률



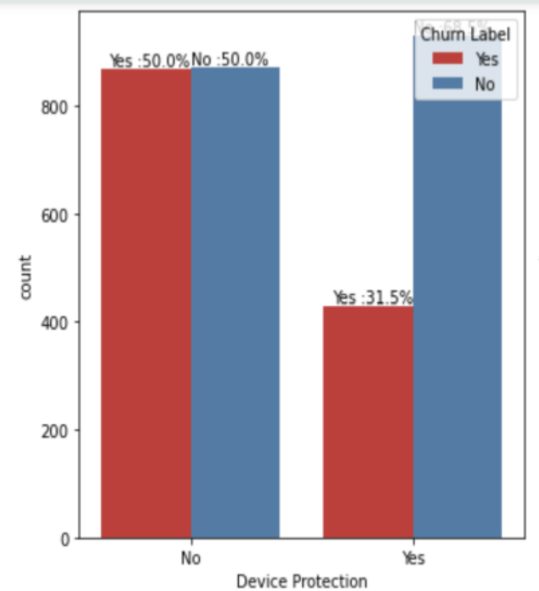
Online Backup



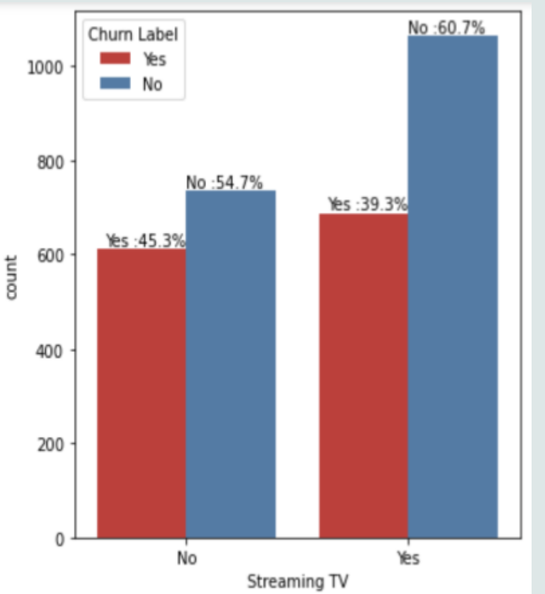
Device Protection



Streaming TV



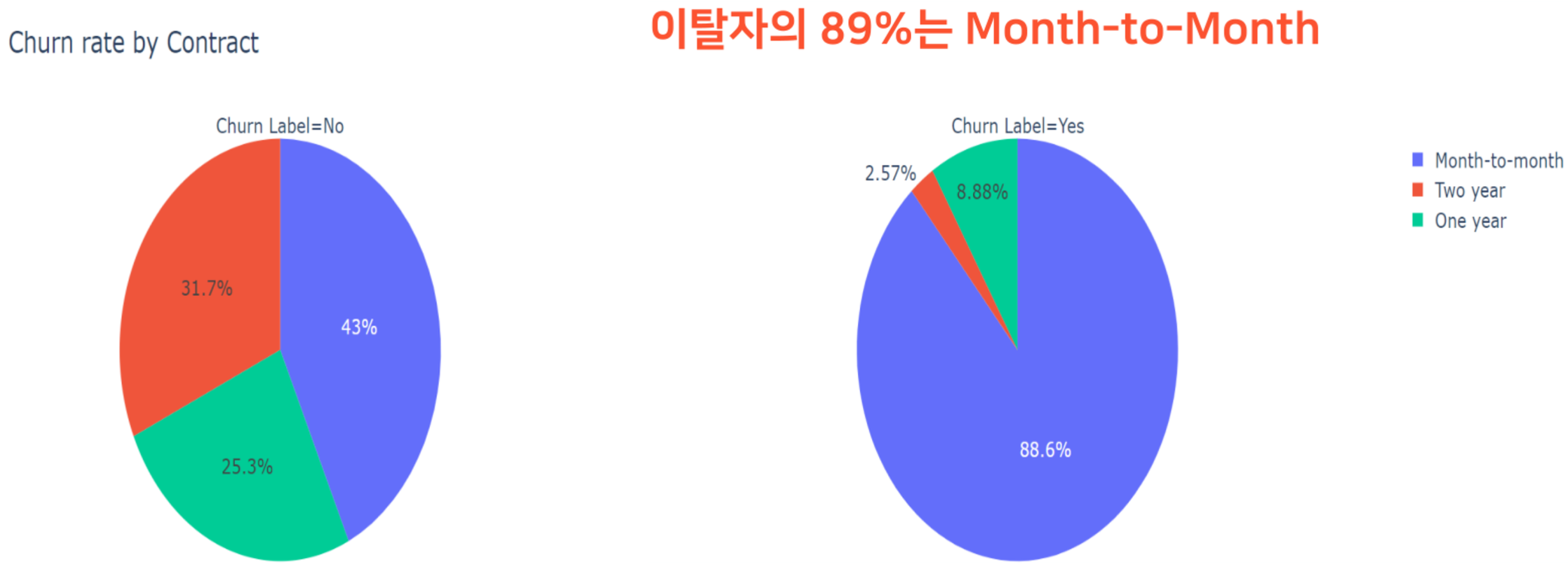
Streaming Movies



Online Security

데이터 EDA - 주요 변수(범주형)

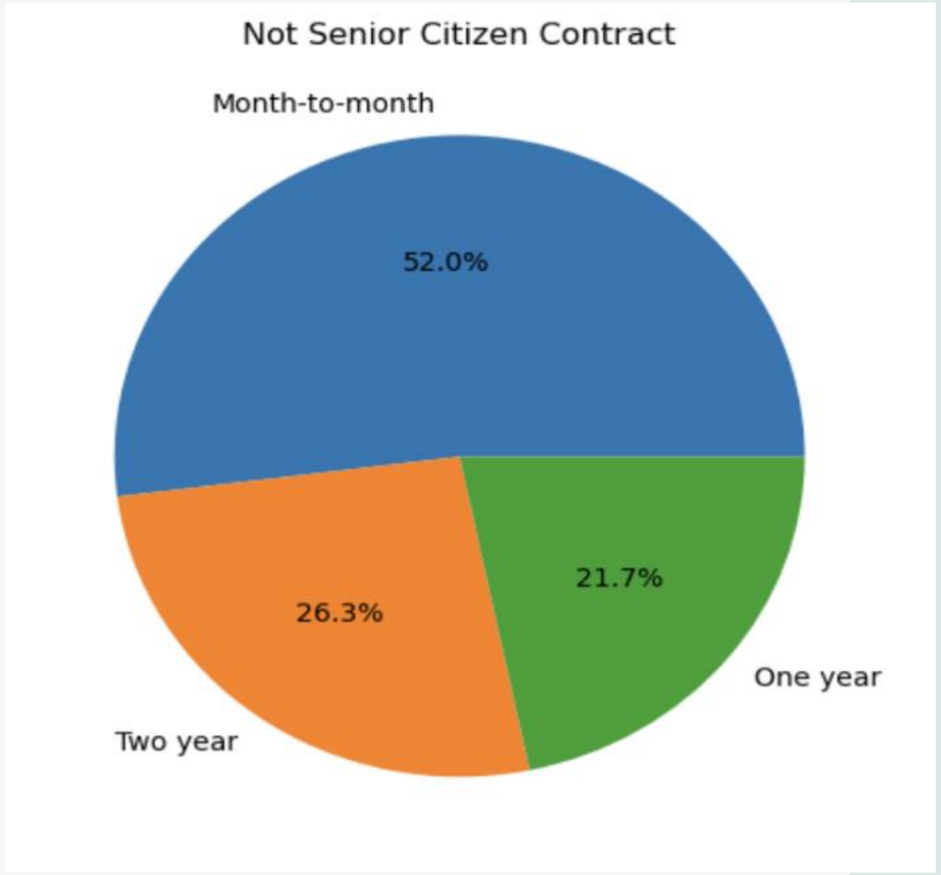
2. Contract

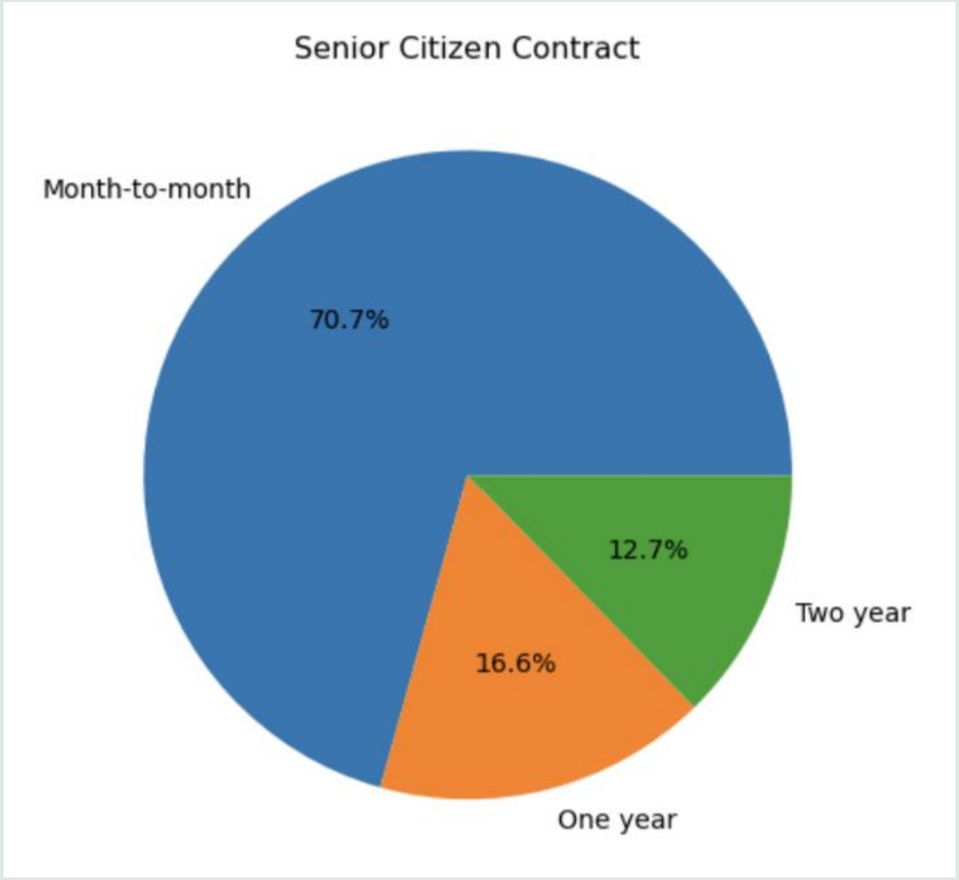


데이터 EDA - 주요 변수(범주형)

2. Contract - Senior Citizen

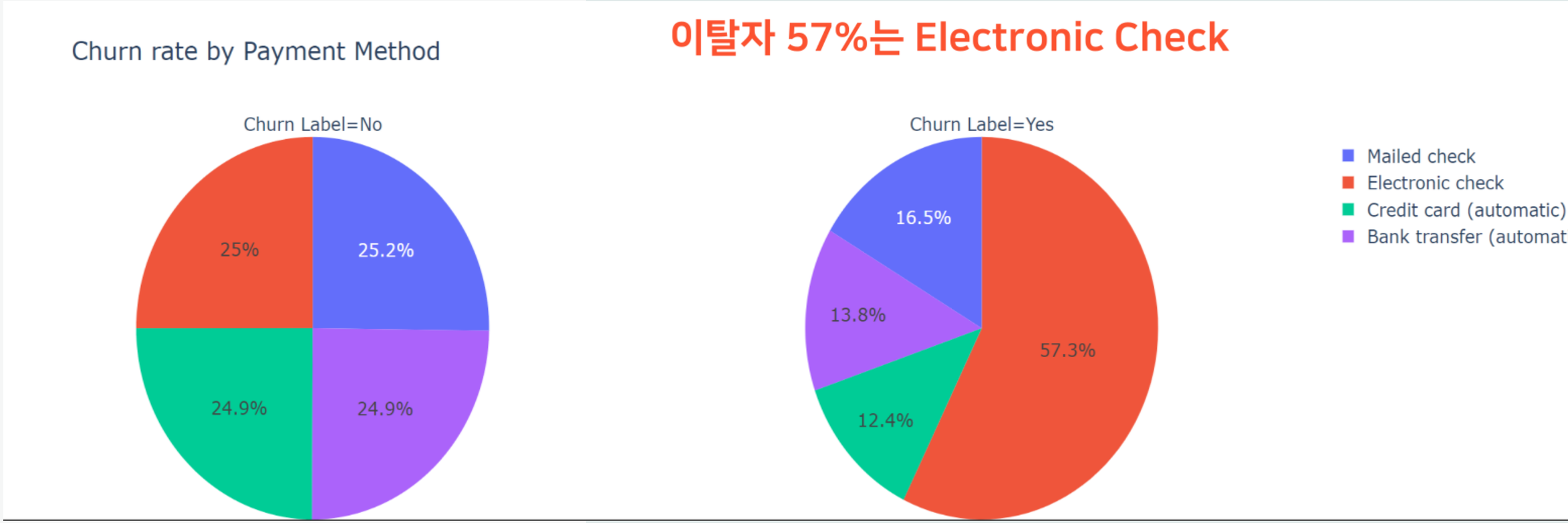
Senior Citizen에 따른 Contract 비율





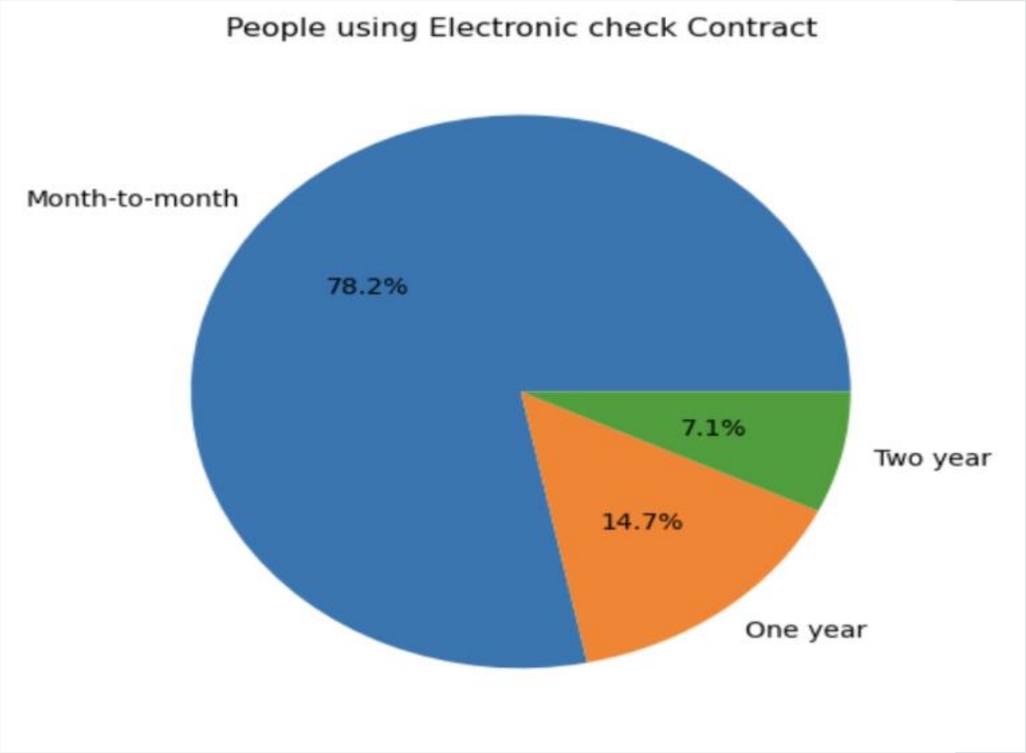
데이터 EDA - 주요 변수(범주형)

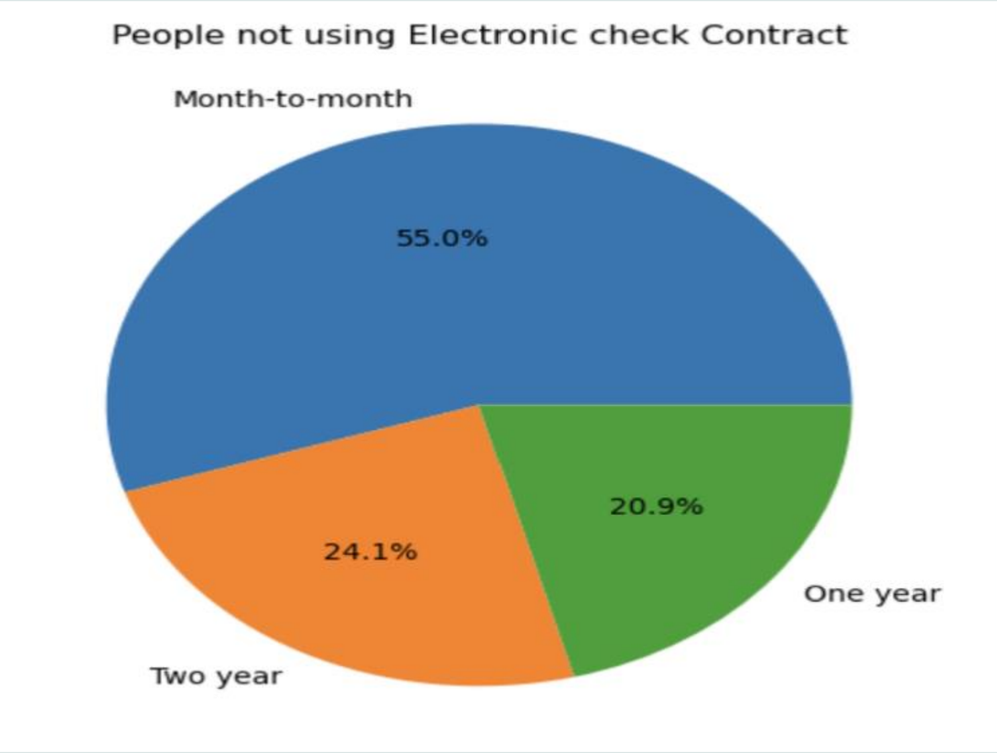
3. Payment Method



데이터 EDA - 주요 변수(범주형)

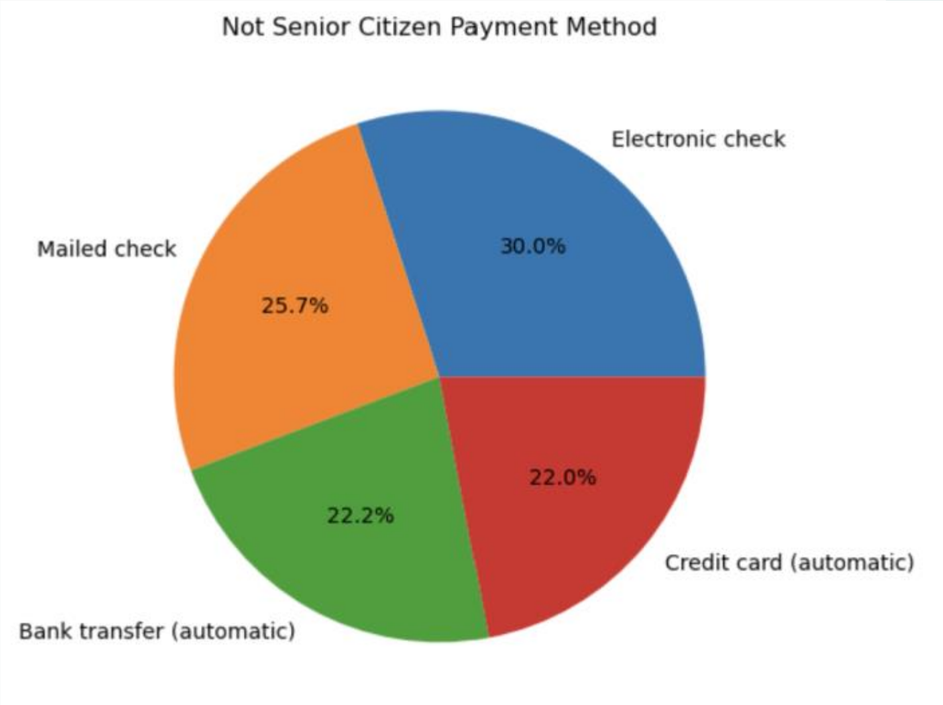
3. Payment Method(Electronic Check) - Contract

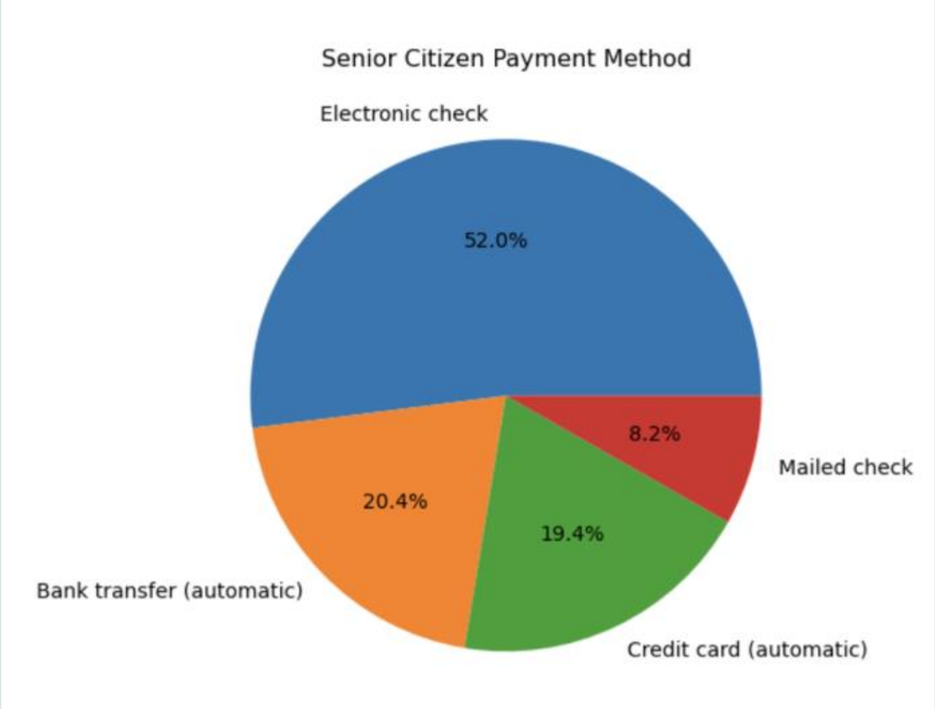




데이터 EDA - 주요 변수(범주형)

3. Payment Method - Senior Citizen





데이터 EDA - 결측치 확인

Numerical 변수가 Categorical로 분류 됨 -> float로 변환

그 중 11개 값은 결측

```
0         108.15
1         151.65
2         820.50
3        3046.05
4        5036.30
          ...
7038     1419.40
7039     1990.50
7040     7362.90
7041      346.45
7042     6844.50
Name: Total Charges, Length: 7043, dtype: float64
```

```
2234     0.0
2438     0.0
2568     0.0
2667     0.0
2856     0.0
4331     0.0
4687     0.0
5104     0.0
5719     0.0
6772     0.0
6840     0.0
Name: Total Charges, dtype: float64
```

데이터 EDA - 결측치 확인

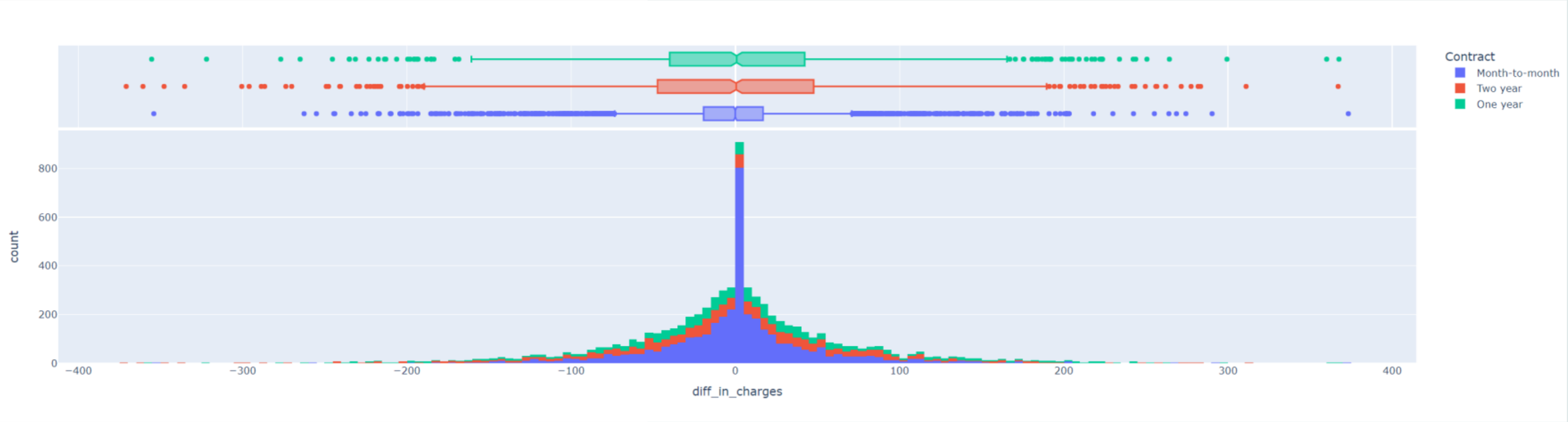
매달 요금 x 이용 기간 = 총 금액

편차를 확인해봄

```
# 결측치 확인
# 총 요금이 (이용 기간 * 달 요금) 값과 편차를 계산. -> 차이를 고려하여 채우기
df[['Monthly Charges', 'Tenure Months', 'Total Charges']]
```

	Monthly Charges	Tenure Months	Total Charges
0	53.85	2	108.15
1	70.70	2	151.65
2	99.65	8	820.5
3	104.80	28	3046.05
4	103.70	49	5036.3
...	...	...	...
7038	21.15	72	1419.4
7039	84.80	24	1990.5
7040	103.20	72	7362.9
7041	29.60	11	346.45
7042	105.65	66	6844.5

7043 rows × 3 columns



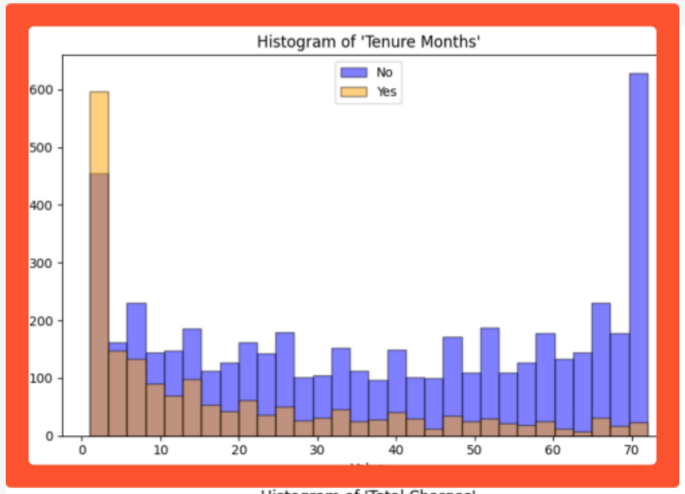
데이터 EDA - 결측치 대체

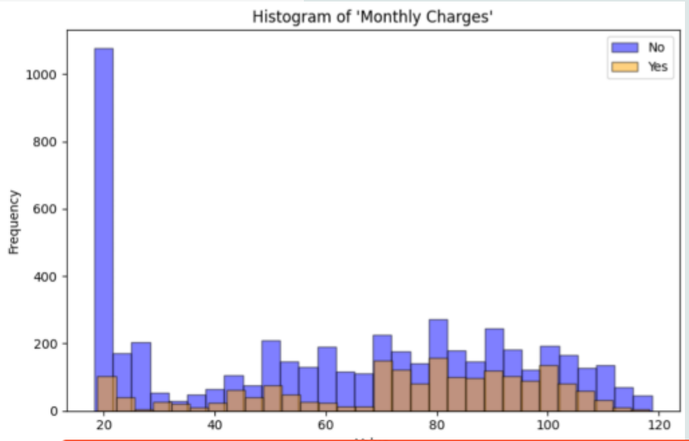
Contract		Total Charges	diff_in_charges
Month-to-month	0.50	679.5500	0.0000
	0.80	2485.7300	24.8100
	0.90	3844.0600	54.0200
	0.95	4966.9200	85.3300
One year	0.50	2657.5500	0.7750
	0.80	5286.4600	55.0500
	0.90	6341.2500	92.2000
	0.95	7072.4725	133.3375
Two year	0.50	3623.9500	0.5000
	0.80	6399.2400	61.5300
	0.90	7457.6100	97.5700
	0.95	7922.3400	139.1800

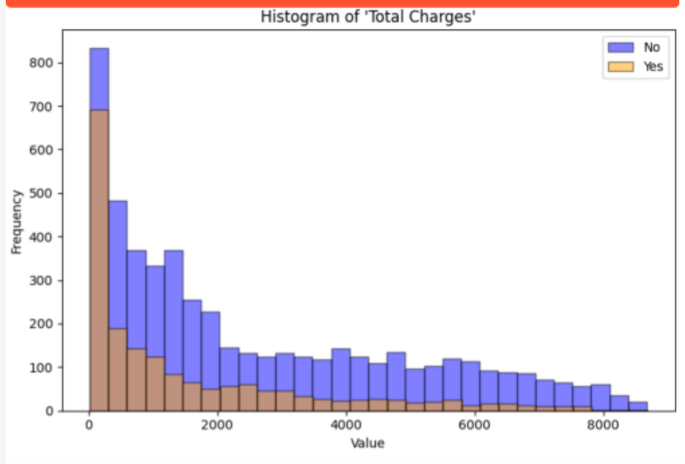
약 20%의 고객이 20 달러 이상의 편차를 가짐

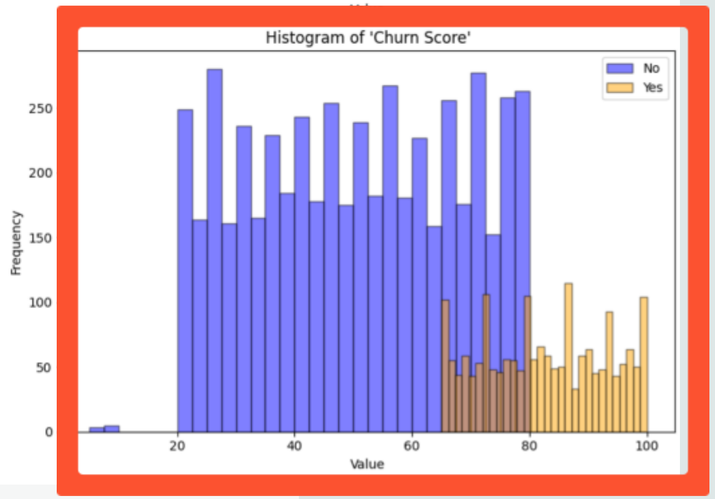
매달 요금 x 이용 기간 = 총 금액 대체

데이터 EDA - 주요 변수(수치형)

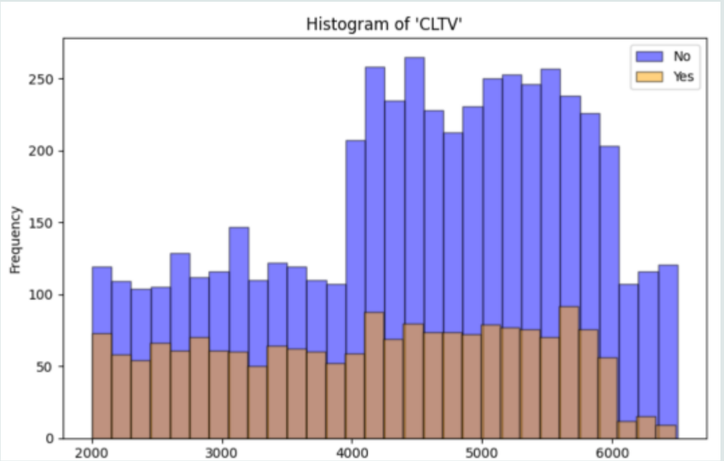




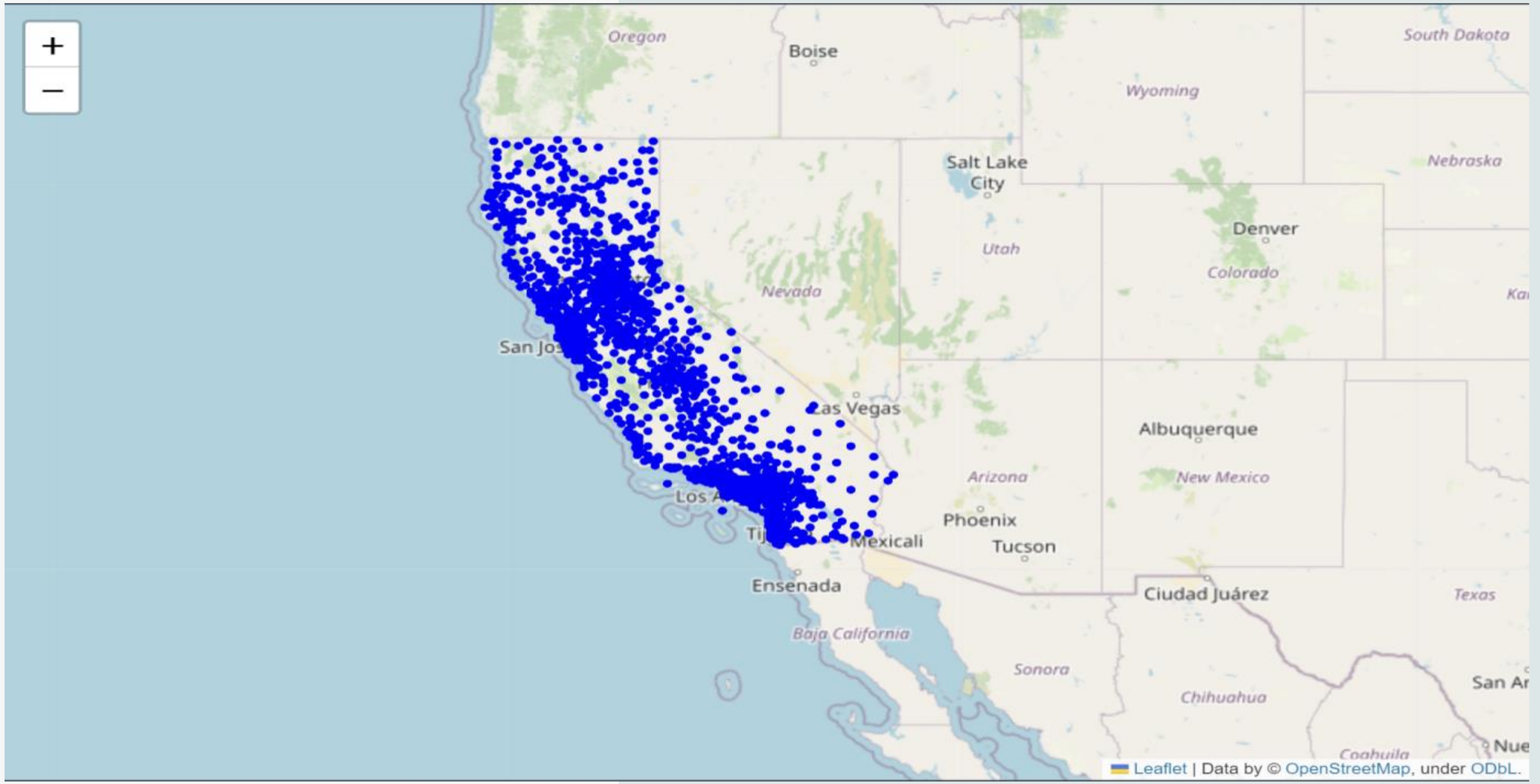




Tenure Months, Churn Score
를 주요 변수로 예상



데이터 EDA - 위도 경도 시각화



데이터 전처리

```
Internet Service :
 Fiber optic     3096
DSL             2416
No              1520
Name: Internet Service, dtype: int64

Online Security :
 No                     3497
Yes                    2015
No internet service    1520
Name: Online Security, dtype: int64

Online Backup :
 No                     3087
Yes                    2425
No internet service    1520
Name: Online Backup, dtype: int64

Device Protection :
 No                     3094
Yes                    2418
No internet service    1520
Name: Device Protection, dtype: int64

Tech Support :
 No                     3472
Yes                    2040
No internet service    1520
Name: Tech Support, dtype: int64

Streaming TV :
 No                     2809
Yes                    2703
No internet service    1520
Name: Streaming TV, dtype: int64

Streaming Movies :
 No                     2781
Yes                    2731
No internet service    1520
Name: Streaming Movies, dtype: int64

Phone Service :
 Yes     6352
No       680
Name: Phone Service, dtype: int64

Multiple Lines :
 No                  3385
Yes                 2967
No phone service     680
Name: Multiple Lines, dtype: int64
```

```python
#필요 없는 열 지우기
df = df.drop(['CustomerID','City','Zip Code','Count','Country','State','Lat Long','Latitude','Longitude',
              'Churn Reason','Churn Value'], axis=1)
```

```python
#Churn 값이 1이면 Yes, 0이면 No이니 다른 열의 값들도 다 바꿔주기
df = df.replace({'Yes': 1, 'No':0 ,'No phone service':0 ,'No internet service': 0})

#Gender의 값들도 바꿔주기
df = df.replace({'Male': 1, 'Female': 0})
```

```python
# unique 값이 적은 컬럼 원핫 인코딩
df = pd.get_dummies(df, columns=['Contract', 'Payment Method','Internet Service'])
df
```

데이터 EDA - Correlation

상관계수 Abs(0.6) 이상 파악

Tenure Months - Contract(Month-to-Month) -0.65

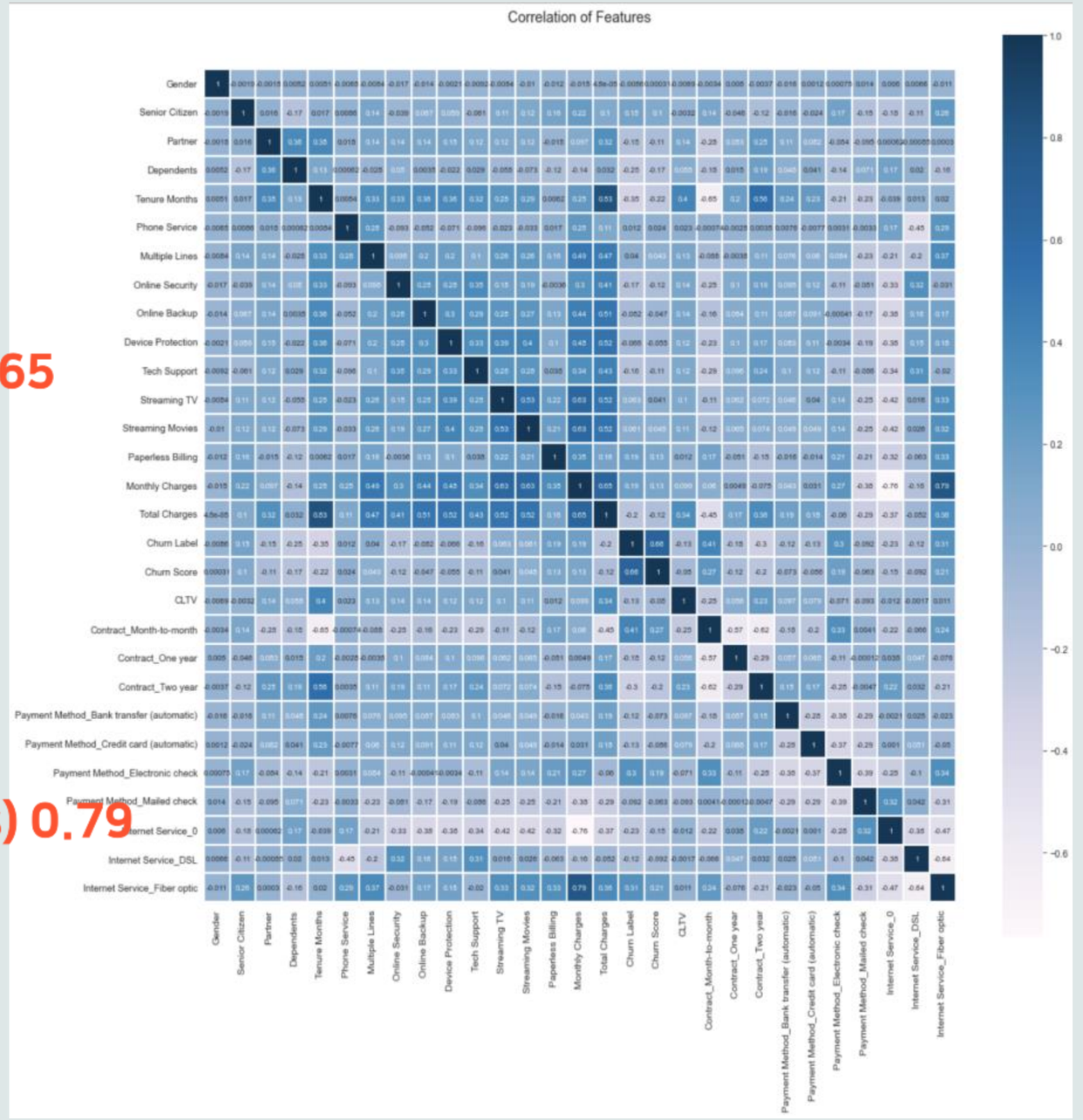
Streaming TV - Monthly Charges 0.63

Streaming Movies - Monthly Charges 0.63

Monthly Charges - Total Charges 0.65

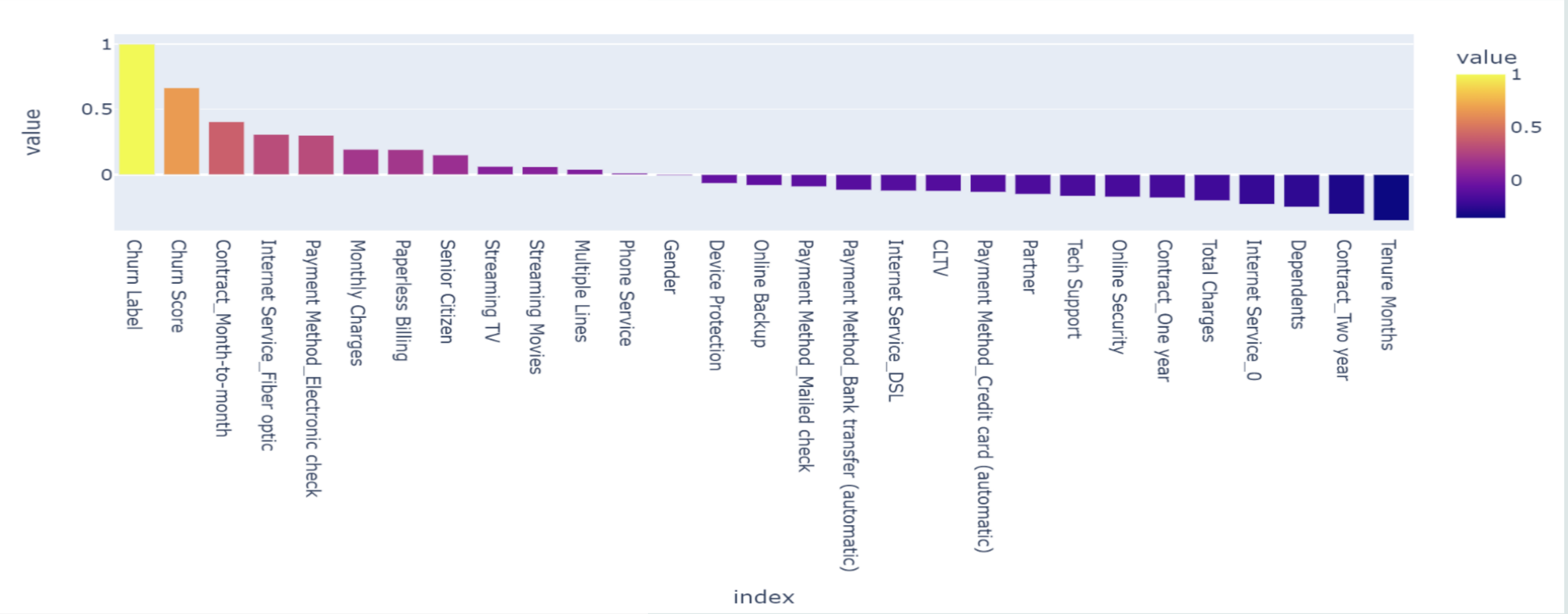
Monthly Charges - Internet Service(Fiber Optics) 0.79

Monthly Charges - Internet Service(No) -0.76

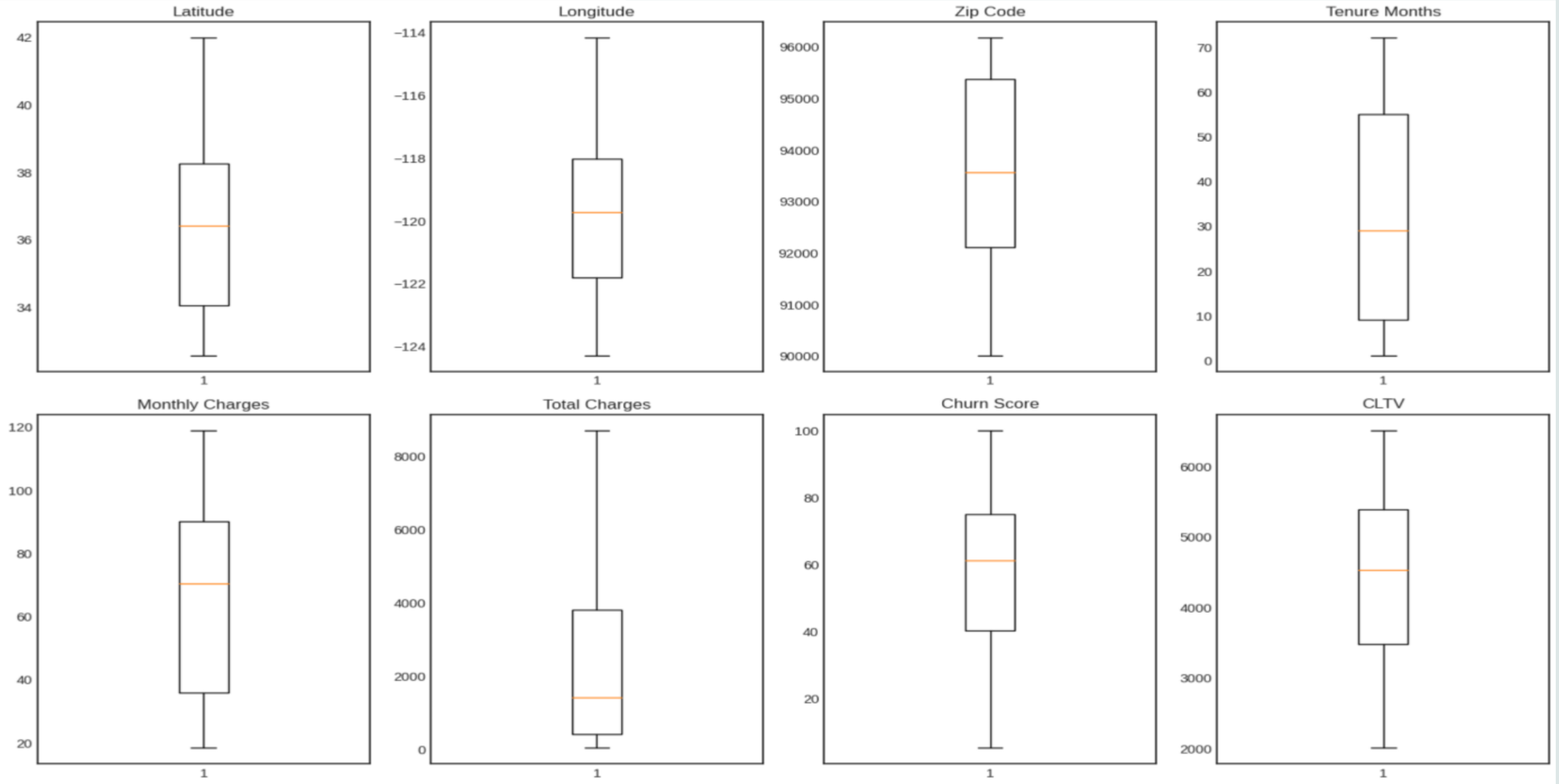


데이터 EDA - Correlation

이탈률과 상관계수가 높은 변수



데이터 EDA - Outlier 확인



수행 계획

월	화	수	목	금	토	일
				5/19	5/20	5/21
				Scaling 및 전처리, 사용 컬럼 추출		
5/22	5/23	5/24	5/25	5/26	5/27	5/28
GB, XGB, LGBM, CatBoost 모델 구축 및 학습						
5/29	5/30	5/31	6/1	6/2	6/3	6/4
GB, XGB, LGBM, CatBoost 모델 구축 및 학습						
6/5	6/6	6/7	6/8	6/9	6/10	6/11
하이퍼 파라미터 최적화						
6/12	6/13	6/14	6/15	6/16	6/17	6/18
최종 결과 보고서 및 발표						

출처

뉴스 기사

https://www.edaily.co.kr/news/read?newsId=02994646635609248&mediaCodeNo=257&OutLnkChk=Y

https://biz.chosun.com/it-science/ict/2022/04/20/2BIRWZY2PNBZJGOAA3FT7E7M2M/?utm_source=naver&utm_medium=original&utm_campaign=biz

https://www.yna.co.kr/view/AKR20220909033100017?input=1195m

Thank You