AR3 불량 분석 용 데이터 셋(new)

▼ **Data Description**

1. 정상데이터 추가

-. 기간 : 2020년 8월 ~ 2020년 10월 (3개월 분)

-. 3개월간 생산량 : 약 397만 개

-. 추출데이터 : 약 330만 건(양품+AR3불량품만 추출, BP1/BP원단/C507/IL/C934 중 하나라도 비어있는 행은 제외)

-. CSV파일로 추출(2.31GB, 오픈 시 별도의 대용량 텍스트에디터 필요)

2. 데이터 필요 기간

-. 3개월 분 추출(2020년 8월~10월)

3. 데이터 셋에 온습도 추가

-. 각 품목 별로 컬럼 추가완료

4. BAD_GRAGE 컬럼에 대한 설명

-. 빈값 : 정상

-. E : 경미불량

-. R1 : 경불량

-. R3 : 중불량

-. K : 판정대기(불량이지만 등급미확정)

-. H : 폐기

-. NT : 폐기(연구소용)

※ BAD_GRADE에 빈값을 제외하고 모두 AR3불량이라 판단하시면 됩니다.

▼ **중요 변수**

CURE_END_TIME : 가류 종료 시간(추정)

TRAN_TIME_0 : 생산시각

INSP_DATE : 검사일자

GT: 그린타이어 (성형 공정 후 만들어짐)

BP : 비드프로세싱 (비드 공정-고무부착)

IL : 이너라이너 (타이어 공기 침투 방지)

Aging Time : 생산-압연(고무를 입히는 과정) 투입, 시간

준비(BP,재단)공정 : 각종 원단류를 타이어 규격에 맞도록 재단하는 고정,

정해진 폭과 각도로 절단된 코드지를 접합하며 라이너에 감아 타이어의 모양을 잡아줄 준비를 하는 과정

▼ 정련 - 압출 - 압연 - 비드(준비) - 성형 - 가류 -완성

1: BP1

2 : BP 원단

3 : BP원단 투입 Compound

4 : IL

5 : IL 투입 compund

▼ **Summary of EDA**

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/c94de085-918d-420d-aa4e-096587dffc78/Exploratory_Data_Analysis.pptx

▼ **종속변수간의 상관관계**

종속변수간 Correlation은 도메인에 대한 확신이 없어 단순한 상관계수는 해석은 의미가

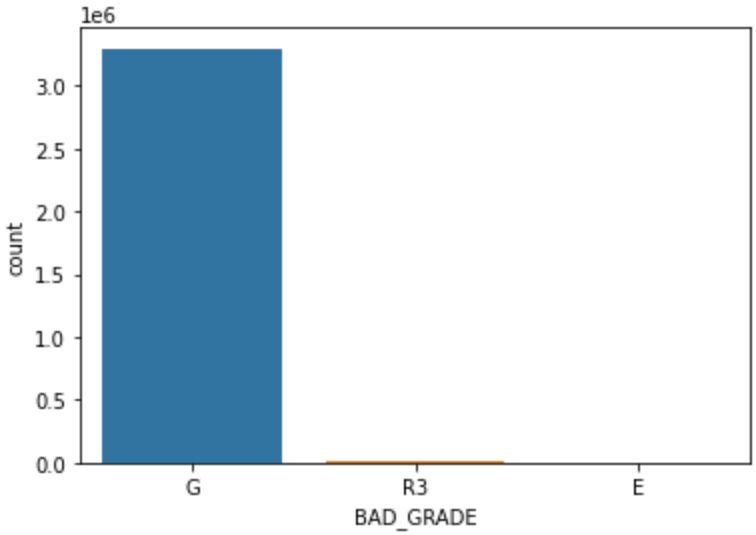
없어보여 진행하지 않음 , 대신 Target 값에 의거한 **조건부 확률**을 통해 변수를 설명

▼ **Bad Grade 관련 시각화**

null 값은 전처리 하기 좋게 → **G**로 치환

→ 대부분의 AR3가 정상이지만 R3와 E가 섞여있음

```
G     3295485
R3       6730
E          19
Name: BAD_GRADE, dtype: int64
```



▼ Bad Grade Imbalanced Class

타겟 값 관련하여 전체 데이터를 본다면 당연히 양품이 많은 **클래스 불균형** 발생

하지만 예측 모델링 진행을 하지 않으므로 Down sampling 등의 추가적인 샘플링은 진행 X

▼ Missing Value 시각화

```
In [354]: pd.set_option('display.max_columns', 240)
          pd.set_option('display.max_rows', 240)
          df.isnull().sum().value_counts().sort_index()

Out[354]: 0         80
          1          4
          19         1
          7294       1
          20012      1
          69450      3
          152011     1
          dtype: int64
```

데이터의 전체적인 missing value를 들여다 보니 가장 많은 null값을 가진 컬럼은 **Series 와Halve_4**

SERIES - 152011

HALB_TEMP_4 - 20012

HALB_TEMP_2 등등

▼ Missing value 처리

공정 결측치는 **도메인적인 파악**이 중요하다

→ 전 회사에서 진행한 병원 환자 **중증도 분류(Triage) 프로젝트**에서 도메인 전문가 의사들과 협업을 했을 때 도메인 개입 없이 단순한 결측치 처리는 **모델 결과에 악 영향**

단순한 결측치 처리를 도메인 없이 **평균이나 KNN값 처리 등은 모델 정확도를 떨어뜨리는 요소로 간주,** 결측치에 대한 원인을 찾는 것에 중점을 두고 진행

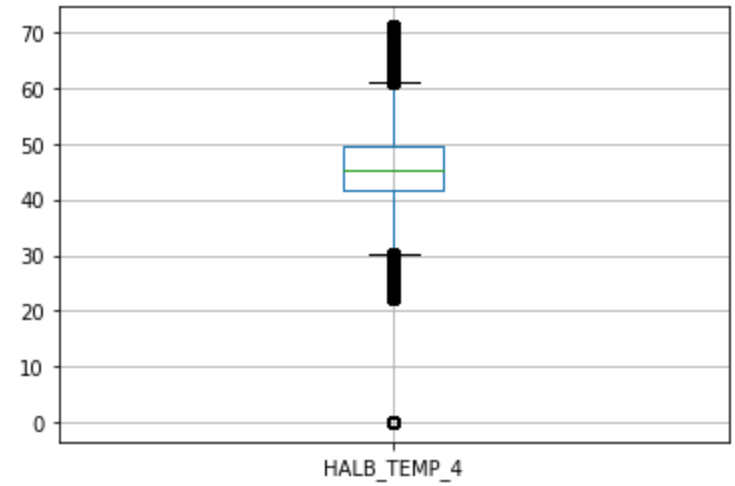
▼ **HALB_TEMP_4 BoxPlot**

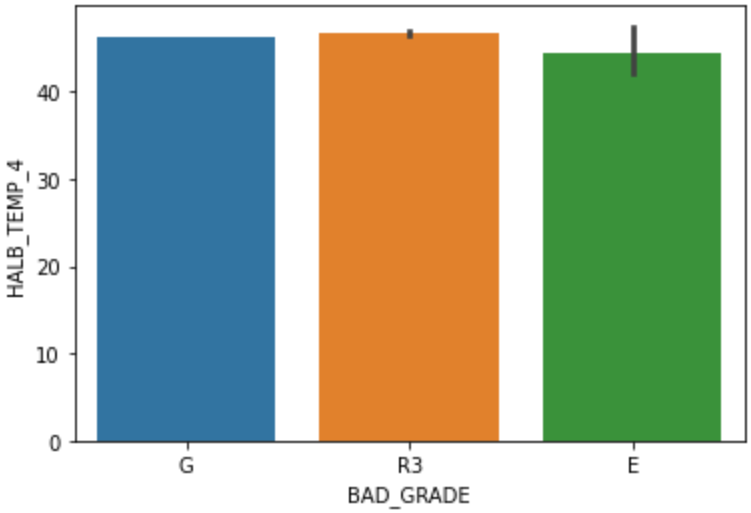
Halb 분포와 Halb에 따른 Bad_grade 분포

압연 과정에서 **반 제품 온도**가 많이 누락 되어있다는 것을 보고 **BAD_GRADE 와의 상관성 확인**

HALB_TEMP4 에 대한 R3 와 E 값이 G값에 비해서도 상당히 많이 존재한다는 것을 파악

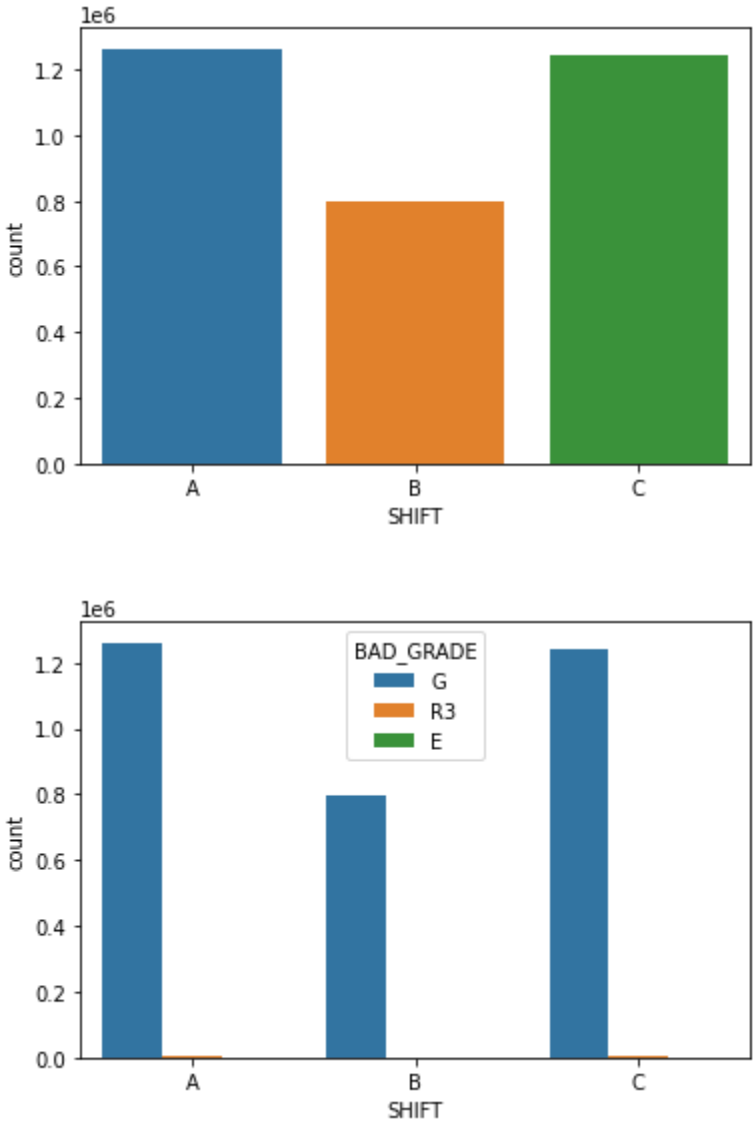
또한 halb는 0값도 많이 가지고 있는데 이것은 결측치에 대한 오 기입으로 간주





▼ **Shift(교대조) 별 시각화**

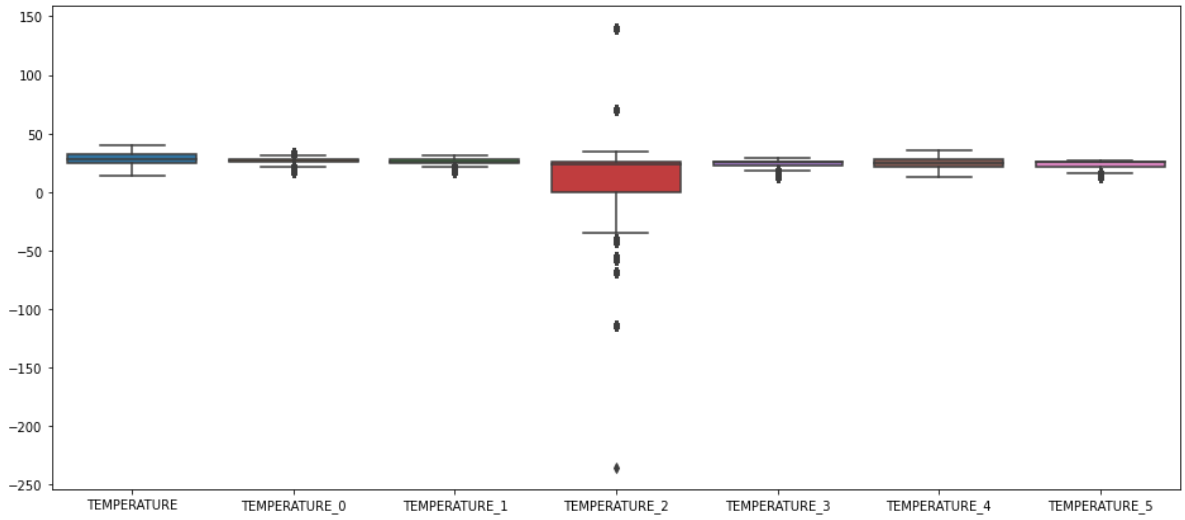
교대조 분포와 교대조에 따른 Bad_grade



▼ **Temperature - 각 공정**

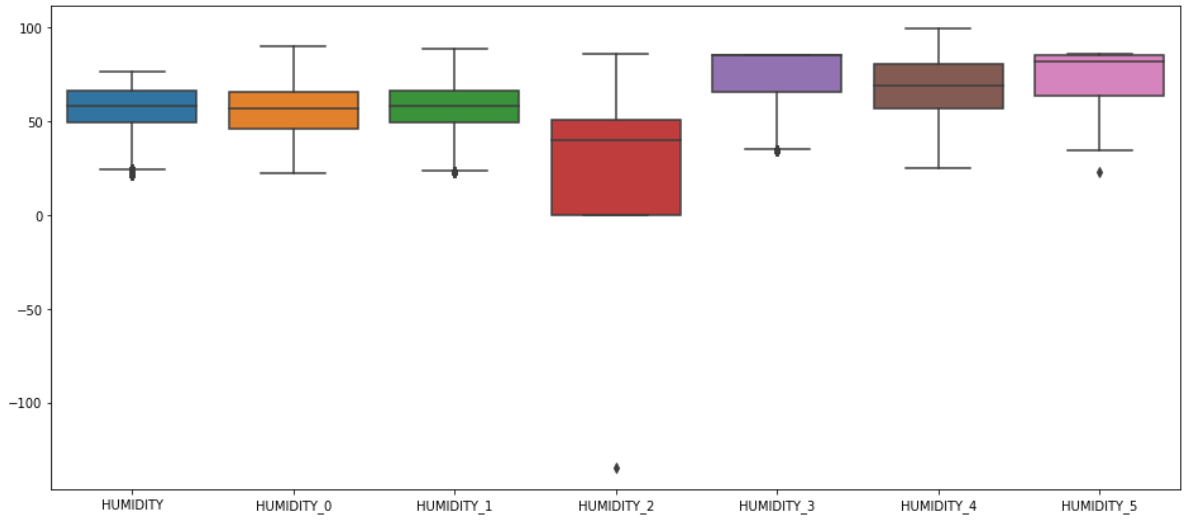
각 공정의 온도는 평균 온도와 비슷했지만

Temperature 2 - 비드프로세싱 압연 작업 때 상 하강 폭이 높은 것을 발견



▼ Humidity - 각 공정

- 각 공정마다 습도의 Boxplot을 확인하면 **Humidity2 비드프로세싱 압연 공정에서 습도가 낮아지는 특이점**을 발견했고 공정 시작부터 완제품으로 갈수록 평균 습도가 낮아짐



▼ 온도 Range별 분포 (Feature Engineering)

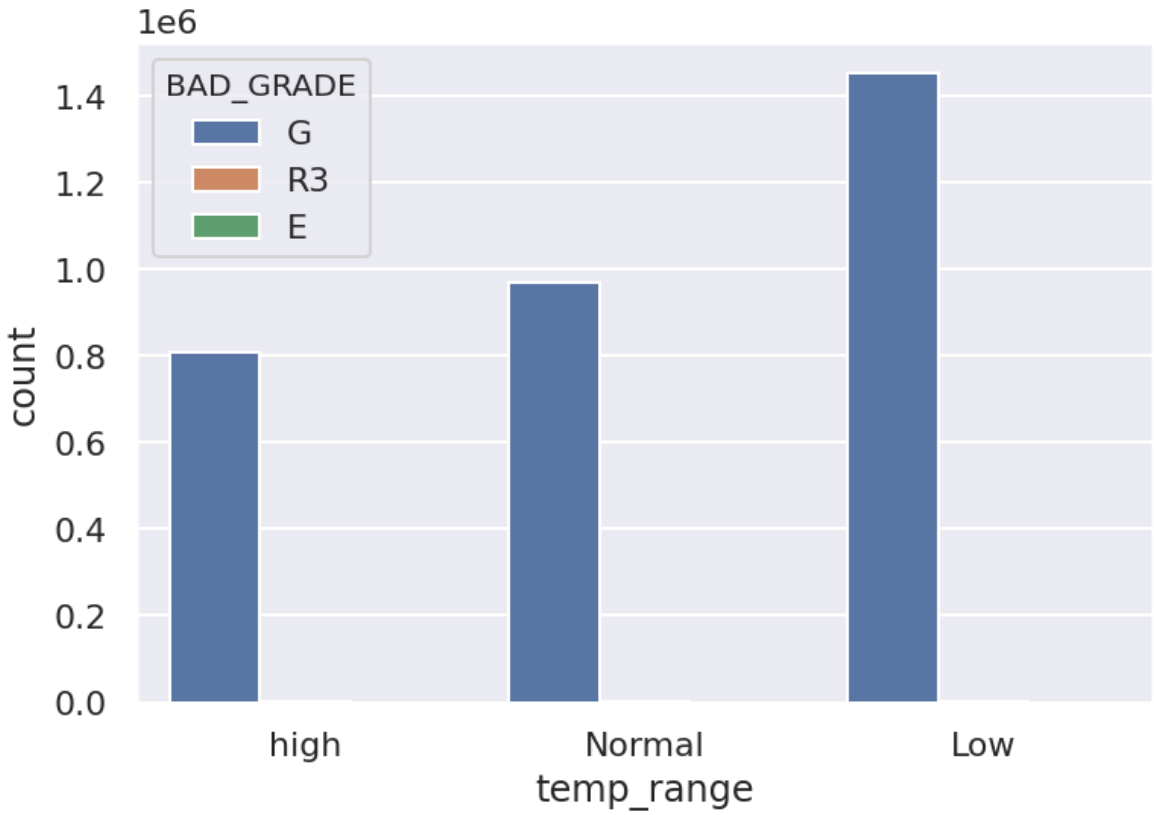
온도의 Quantile 살펴보고 25% 50% 75%별로 연속 형 변수를

직관성을 높이고 모델링 편의성을 높이기 위해→ **범주형 변수로 바꾸고 Range화를 진행**

온도 분포 별 Range 정의

```
round(df['TEMPERATURE'].describe(),2)
```

```
count    3232784.00
mean          27.83
std            5.09
min           14.05
25%           24.29
50%           27.71
75%           32.02
max           40.25
Name: TEMPERATURE, dtype: float64
```



▼ 작업자 분포

공정의 작업자 별로 불량률을과 빈도수를 살펴 봤는데 큰 의미를 찾지 못함

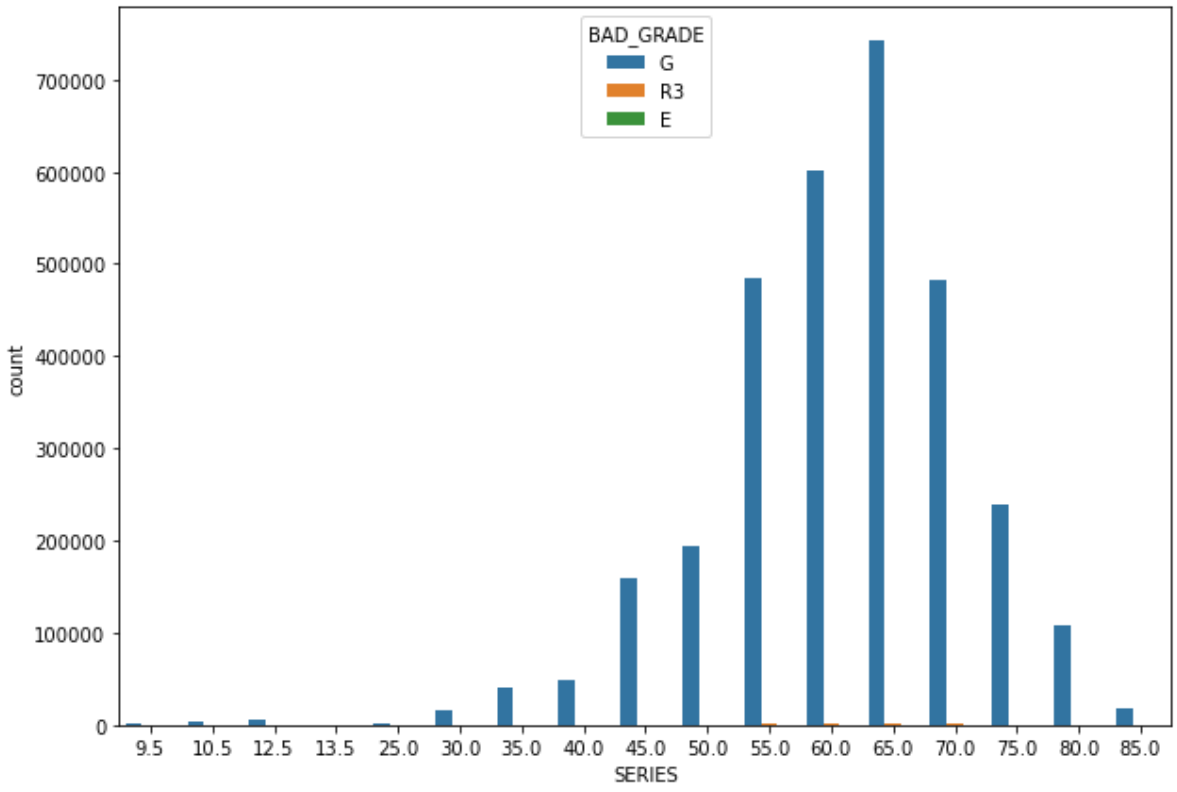
```
df['USER_DESC_5'].value_counts().head()
```

```
정지원    757244
김종문    467243
이정근    322298
정임도    208610
김남진    176497
Name: USER_DESC_5, dtype: int64
```

▼ **Series 분포**

완제품 시리즈 에 따른 BAD_GRADE 분포 - 거의 G 상태인 것으로 보임

→ 시리즈를 양품과 불량을 **상대도수로 히스토그램으로 표현하기**



R3 상태인 시리즈도 적지만 몇 개씩 볼 수 있으므로 불량률을 Series에 따라 분류하는 것도 괜찮아 보임

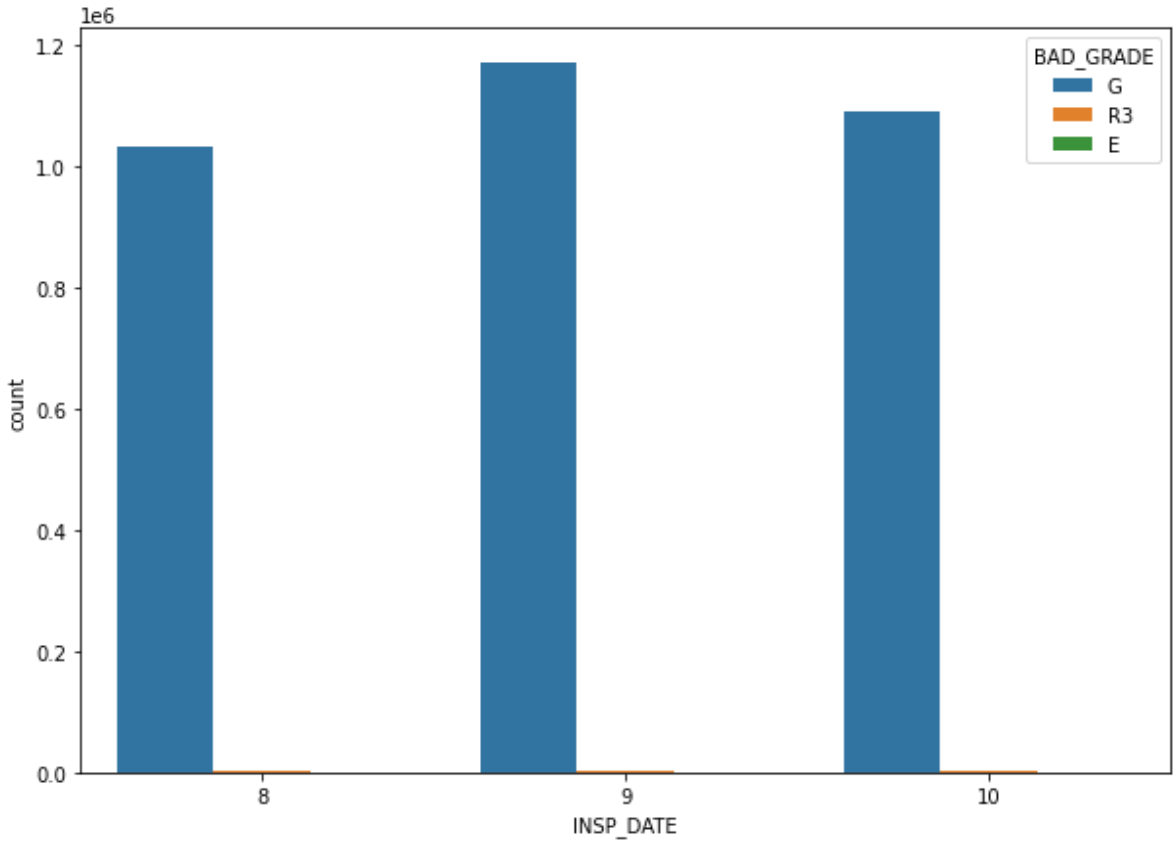
```
In [262]: df.loc[df['BAD_GRADE']=='R3','SERIES']

Out[262]: 331        60.0
          344        70.0
          580        65.0
          624        70.0
          625        70.0
                     ...
          3296362    35.0
          3296902    45.0
          3296928    40.0
          3296961    40.0
          3301071    60.0
          Name: SERIES, Length: 6730, dtype: float64
```

▼ **검사 월별**

완 제품 검사일자에 대해 월 별 불량률을 알아보기 위해 시각화 진행

→ 8월이 불량률이 전체적으로 높다는 것을 볼 수 있고 E값도 높은 것으로 보아

온도가 높은 여름에 대한 인사이트를 찾아 볼수있음




```
df[df['BAD_GRADE']=='R3']['INSP_DATE'].dt.month.value_counts()
```

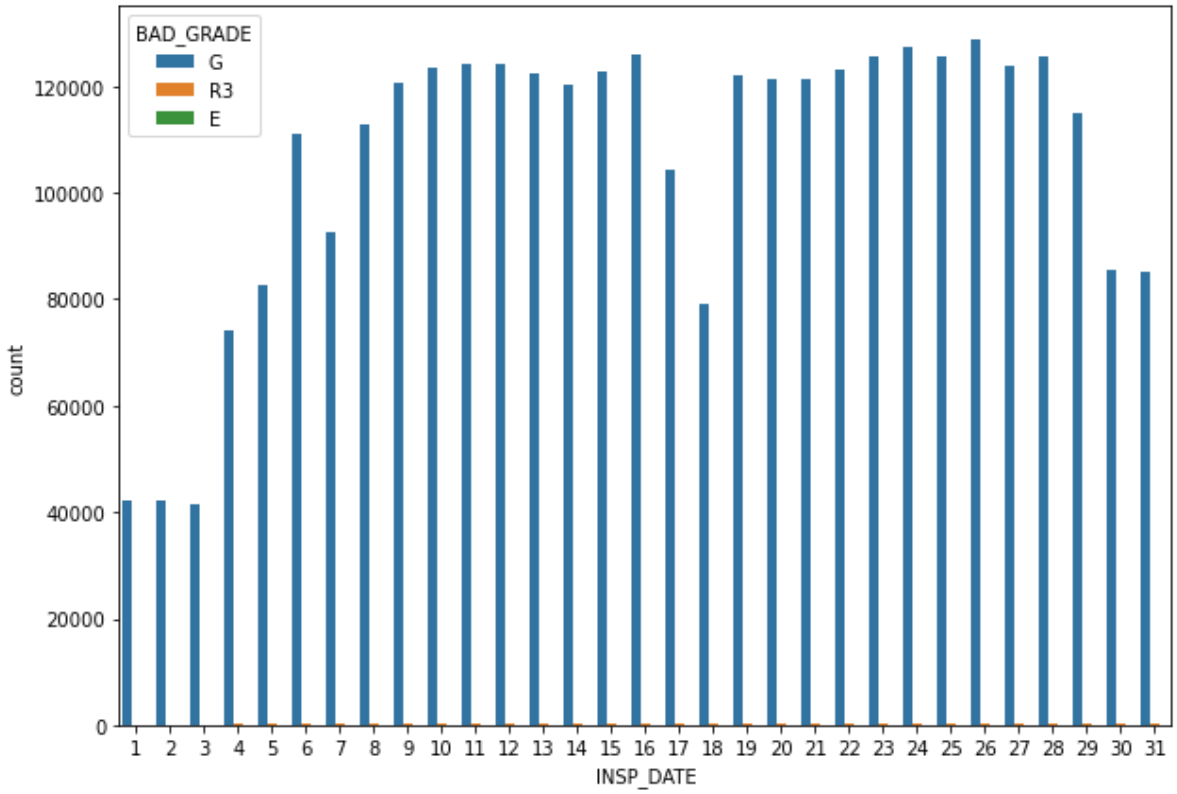
```
8     2821
9     2532
10    1377
Name: INSP_DATE, dtype: int64
```

```
df[df['BAD_GRADE']=='E']['INSP_DATE'].dt.month.value_counts()
```

```
8     11
9      7
10     1
Name: INSP_DATE, dtype: int64
```

▼ 검사 일별

매월 초와 매월 중순 , 말 즈음에 양호한 타이어의 갯수가 낮아지는 것을 볼 수 있음



▼ **결론**

Focusing 을 해야하는 부분은 온 습도와 Aging Time

외부 데이터를 지역 별로 (양산) 연동해서 어떤 다양한 Insight를 얻을 수 있을지 파악해 보기

일별 Heat Map 확인하기