

데이터분석 포트폴리오



데이터분석가 유 정 호

## Ideation



동일한 비즈니스 프로세스 : 차량통과 중 - 결제, (상품)전달, (니즈)달성

고속도로 유희시설을 이용하여 휴게소 불만(비싸고 맛있는 커피) 해결 방안 마련

## Ideation



유희시설(톨게이트 요금소) 철거계획 확인



시설물 관리주체(한국도로공사)의 사업의지 확인

사업참여의지가 있는 커피 프랜차이즈 모집

톨게이트 요금소를 이용한 커피전문점 영업의 구체적 방침 수립



가장 먼저 할 수 있는 예비타당성 검토 (매몰비용 ≒ 0)



핵심 비즈니스 활동 (관리주체 승인 없이 사업시행 불가능)

## Ideation

### 유희시설(톨게이트 요금소) 철거계획 확인

김학송 도공 사장 "2020년 '스마트 톨링'...톨게이트 없앤다"

"하이패스 보급률이 80% 넘어가면 톨게이트를 전부 다 없애고 그냥 지나가면 되는데 그때가 2020년쯤 될 것 "



- 2020년이 훨씬 넘어 2024년 일 뿐 아니라, 하이패스 보급율과 실제 통행권을 이용하는 톨게이트 이용 차량들의 통행량 사이에는 상관성이 다소 부족
- 도로공사의 톨게이트 철거 의지는 현재진행 상태(스마트 톨링 사업 단계적 진행)

🔑 실제 통행권을 사용하는 톨게이트 통행량 증감 현황에 대한 분석 필요

# Data Analysis

- Dataset**

한국도로공사 고속도로 공공데이터 포털(톨게이트별 일일 교통량 데이터셋)

<http://data.ex.co.kr/portal/fdwn/view?type=TCS&num=C7&requestfrom=dataset#>

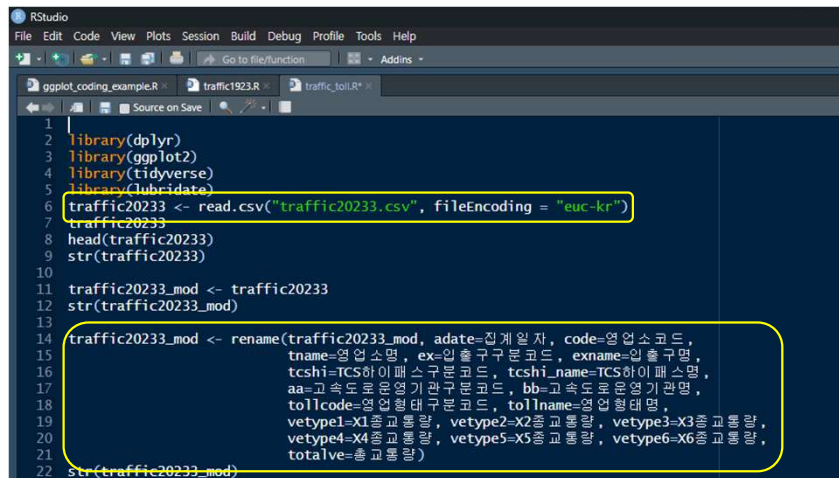
- 원본 데이터셋 구조

#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	집계일자	영업소코드	영업소명	입출구구분코드	입출구명	TCS하이패스구분코드	TCS하이패스명	고속도로운영기관구분	고속도로운영기관명	영업형태구분코드	영업형태명	1종교통량	2종교통량	3종교통량	4종교통량	5종교통량	6종교통량	총교통량
2	2023-07-01	246	가락	0	입구		1 TCS	0	한국도로공사		0 폐쇄식	157	6	40	30	10	2	245
3	2023-07-01	246	가락		0	입구	2 hi-pass		0 한국도로공사		0 폐쇄식	572	24	35	49	94	26	800
4	2023-07-01	29	가락(개)		0	입구	1 TCS		0 한국도로공사		1 개방식	1518	26	68	35	77	198	1922
5	2023-07-01	29	가락(개)		0	입구	2 hi-pass		0 한국도로공사		1 개방식	12139	121	142	101	813	674	13990

**18개 변수(variables)구성, 분기별로 데이터셋 구분(예; 2023년 3분기), csv형식**

## Preprocessing

- 데이터 프레임내 한글 인식을 위한 인코딩 및 변수명 변경, 그리고 해석을 위한 데이터셋 준비



```
1 |
2 | library(dplyr)
3 | library(ggplot2)
4 | library(tidyverse)
5 | library(lubridate)
6 | traffic20233 <- read.csv("traffic20233.csv", fileEncoding = "euc-kr")
7 | traffic20233
8 | head(traffic20233)
9 | str(traffic20233)
10 |
11 | traffic20233_mod <- traffic20233
12 | str(traffic20233_mod)
13 |
14 | traffic20233_mod <- rename(traffic20233_mod, adate=집계일자, code=영업소코드,
15 |                           tname=영업소명, ex=인출구구분코드, exname=인출구명,
16 |                           tcshi=TCS하이패스구분코드, tcshi_name=TCS하이패스명,
17 |                           aa=고속도로운영기관구분코드, bb=고속도로운영기관명,
18 |                           tollcode=영업형태구분코드, tollname=영업형태명,
19 |                           vetype1=X1종교통량, vetype2=X2종교통량, vetype3=X3종교통량,
20 |                           vetype4=X4종교통량, vetype5=X5종교통량, vetype6=X6종교통량,
21 |                           totalve=총교통량)
22 | str(traffic20233_mod)
```

- R(studio)**에서 한글 인식이 안되는 경우, 대부분의 지침 및 가이드에서는 **Default text encoding: UTF-8**로 변경 권고
- Stack overflow** 질의 응답을 조사해 본 결과,  
**Default text encoding: euc-kr**, 변수명은 영어로 변경한다는 의견이 지배적 – 반영하여 **R-script** 작성

- 분기별로 나누어져 있는 데이터프레임 하나로 합치기: 2019년 1월1일 부터 2023년 9월 30일까지
- 해석준비가 완료된 데이터 셋은 csv와 함께 R 포맷인 .rds로 백업

# 소스코드 - [http://www.github.com/AndersonAt17/R\\_data\\_science](http://www.github.com/AndersonAt17/R_data_science)

# 원본데이터 - [https://drive.google.com/drive/folders/11rRa5MK5dSv6EQtFTAafuFI4j\\_-euMG](https://drive.google.com/drive/folders/11rRa5MK5dSv6EQtFTAafuFI4j_-euMG)



- 데이터 전처리 및 EDA

- 요금소 속성별(하이패스, 통행권 요금소; tcs) 데이터 프레임 분리
- 데이터 프레임 변수 조정 및 통합: 1, 6종 차량: 승용차(car), 2~5종 차량: 화물차(truck)
- 데이터 결측치(missing value) 및 이상치(outlier) 탐색 및 조치(영업소별, 요금소 속성별 분류)

```
#-----1. tcs 서울(code = 101, car), 이상치(outlier) 확인
traffic1923_tcs_101 <- traffic1923_tcs %>%
  filter(code == 101 | tname == '서울') %>%
  select(adate, code, tname, car, truck)
head(traffic1923_tcs_101)
qplot(traffic1923_tcs_101$car, geom="boxplot")
qplot(traffic1923_tcs_101$truck, geom="boxplot")
#-----1. tcs 서울(code = 101) 이상치(outlier) 제거
traffic1923_tcs_101$car <- ifelse(traffic1923_tcs_101$car >= 17500, NA, traffic1923_tcs_101$car)
traffic1923_tcs_101$truck <- ifelse(traffic1923_tcs_101$truck >= 3000, NA, traffic1923_tcs_101$truck)
str(traffic1923_tcs_101)
view(traffic1923_tcs_101)
```

# 소스코드 - [http://www.github.com/AndersonAt17/R\\_data\\_science](http://www.github.com/AndersonAt17/R_data_science)

# 원본데이터 - [https://drive.google.com/drive/folders/11rRa5MK5dSv6EQtFTAafuFI4j\\_-euMG](https://drive.google.com/drive/folders/11rRa5MK5dSv6EQtFTAafuFI4j_-euMG)

## Data Analysis and Visualization

- 데이터 분석 및 시각화

- 1477개 영업소 중 교통량 상위 10개소 추출 및 Bar chart 작성
- 교통량 상위 10개소 승용차/화물차로 구분하여 데이터 시각화(scatter plot) 코드 작성
- 교통량 1위 영업소(서울 톨게이트) 회귀분석 및 결제방식별 교통량 증감 추이 분석

```
129
130 #-----1. tcs 서울(code = 101) 교통량 차트작성: 승용차
131 Base_plot_101_car <- ggplot(data = traffic1923_tcs_101_ratio, aes(x=adate, y=car, color = traffic1923_tcs_101_ratio$car)) +
132   geom_jitter() + geom_smooth(method = 'loess', color = 'yellow', span = 2)
133
134 Addon_title_Base101_car <- Base_plot_101_car + ggtitle("Traffic Variation Pass Through \n TCS during 2019_2023: Car") +
135   theme(plot.title = element_text(family = "serif", face = "bold", hjust = 0.5, size = 15, color = "darkblue"))
136
137 Addon_legend_Base101_car <- Addon_title_Base101_car + theme(legend.position = "none") # 범례 관련 편집_현재는 범례 삭제
138
139 Addon_legend_Base101_car
140
141 ggsave("Plot_101_tcs.jpg", dpi = 300)
142
```

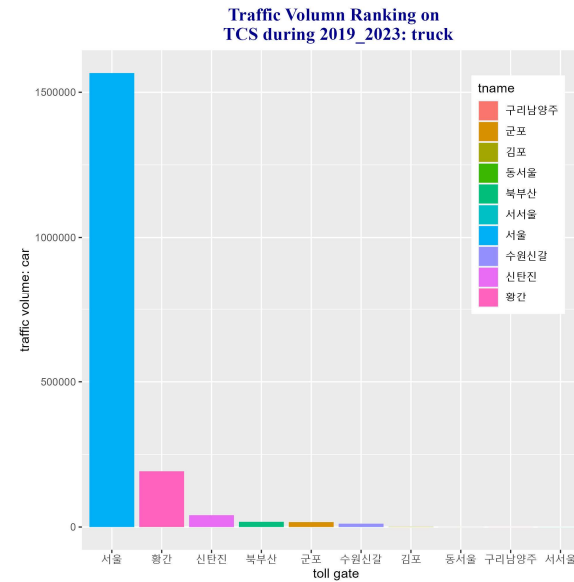
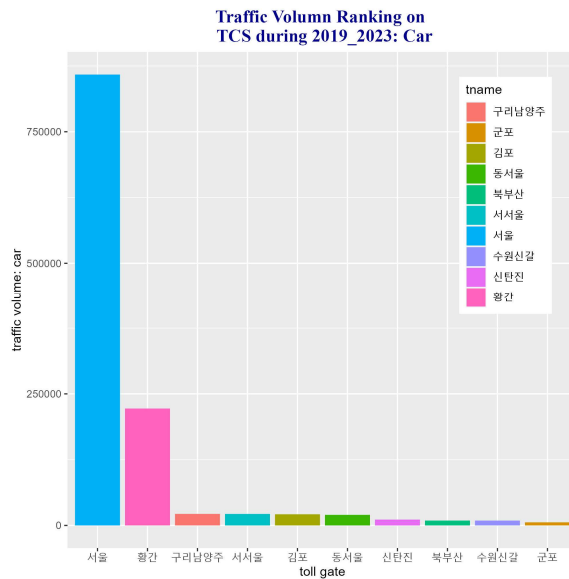
# 소스코드 - [http://www.github.com/AndersonAt17/R\\_data\\_science](http://www.github.com/AndersonAt17/R_data_science)

# 원본데이터 - [https://drive.google.com/drive/folders/11rRa5MK5dSv6EQtFTAafuFI4j\\_-euMG](https://drive.google.com/drive/folders/11rRa5MK5dSv6EQtFTAafuFI4j_-euMG)



## Data Analysis and Visualization

- 데이터 시각화 및 분석 결과



- tcs 분야에서 일일통행량이 가장 많은 영업소: 서울 톨게이트

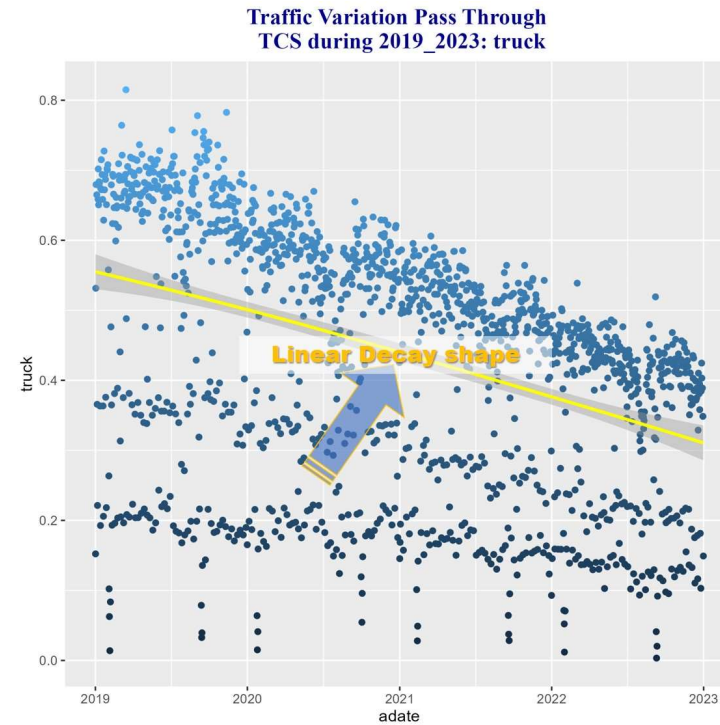
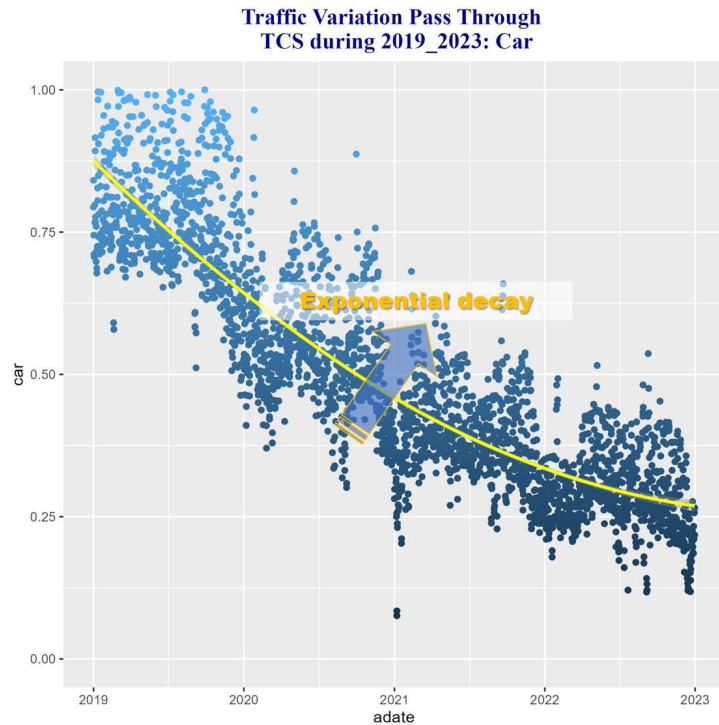
- 서울 톨게이트 대상 tcs 통행량을 승용차 그룹과 화물차 그룹에 대해 각각 시각화 및 분석

# 소스코드 - [http://www.github.com/AndersonAt17/R\\_data\\_science](http://www.github.com/AndersonAt17/R_data_science)

# 원본데이터 - [https://drive.google.com/drive/folders/11rRa5MK5dSv6EQtFTAafuFI4j\\_-euMG](https://drive.google.com/drive/folders/11rRa5MK5dSv6EQtFTAafuFI4j_-euMG)

## Data Analysis and Visualization

- 데이터 시각화 및 분석 결과



- 서울 톨게이트에서 승용차 그룹과 화물차 그룹이 각각 exponential decay, linear decay 경향을 보임  
: 지속적인 감소경향을 보이지만 급속한 감소없이 일정한 비율로 계속 유지될 것으로 분석됨

# 소스코드 - [http://www.github.com/AndersonAt17/R\\_data\\_science](http://www.github.com/AndersonAt17/R_data_science)

# 원본데이터 - [https://drive.google.com/drive/folders/11rRa5MK5dSv6EQtFTAafuFI4j\\_-euMG](https://drive.google.com/drive/folders/11rRa5MK5dSv6EQtFTAafuFI4j_-euMG)

## Data Analysis and Visualization

- 본 분석의 한계 및 보완사항

- “Exponential decay”에 대해 정밀한 회귀분석 필요: coefficient 산출 및 예측 모델 구성

- 분석 시작시점(2019년 1월 1일)을 기준으로 지속적으로 업데이트를 통해 예측모델과 관측결과 비교 분석