1.상관관계 분석

주제 : 대한민국 도로운송 부문에서 화물차 적재율과 탄소배출 효율성의 상관관계 분석

연구가설(H1) : 화물차 적재율과 단위운송거리당(ton*km) 탄소배출량은 상관관계가 있다.

적재율 = 실제 운송톤수/(차량 적재가능톤수*차량대수)

탄소배출량 = 평균 운행 거리*차종별 배출계수

사용 데이터 :

- 우리나라 총 운송량 데이터(평균 적재율로 상관관계만 파악하고자 하므로 운송수단 구분 필요 없이 총 합만 구하면 됨,실제 운송톤수 추정 가능)

  국가물류통합정보센터(NLIC)-내륙화물통계(국가물류통합정보센터)

  단위 : 톤

목적지 / 출발지	서울
서울	43,496,606
부산	1,438,857

- 우리나라 운송수단 등록 현황 데이터(적재가능 톤수와 차량대수 추정가능)

  국가물류통합정보센터(NLIC)-운송수단통계-적재량별 화물차등록현황(국가물류통합정보센터

구분		
	1월	2월
1톤미만	3,005,076	3,003,957
20톤이상	110,540	110,664
1톤~3톤	225,925	225,434

- 한국교통안전공단(CTIS) 차종별 탄소배출계수(downloadBbsFile.do)

  탄소배출량 계산에만 쓰이는 데이터

화물	소형	경유	65.4km/h 미만	$y = 1250.4831x^{-0.4630}$
			65.4km/h 이상	$y = 0.0292x^{2} - 2.9530x + 258.3205$
	중형	경유	64.7km/h 미만	$y = 1385.8860x^{-0.4184}$
			64.7km/h 이상	$y = 1.6720x + 141.2224$
	대형	경유	$y = 3351.2892x^{-0.4407}$	

해당 주제 특징 : 전처리와 새로운 데이터를 생성하는 과정을 강조할 수 있음,물류분야 인사이트 강조 가능,평균 운행 거리를 50km상수값으로 고정하거나 적절하게 처리하는 방법 찾아야함

2.확률분포 분석

주제 : 항공편 일별 지연 건수의 확률분포 분석

연구가설(H1) : 항공편의 일별 지연 건수는 포아송분포를 따를 것이다.

사용 데이터 : DACON 항공편 지연 데이터 (월간 데이콘 항공편 지연 예측 AI 경진대회 - DACON)

ID	Month	Day_of_M	Estimated_	Estimated_	Cancelled	Diverted	Origin_Air	Origin_Air	Origin_Sta	Destinatio	Destinatio	Destinatio	Distance	Airline	Carrier_Co	Carrier_ID(	Tail_Numb	Delay
TRAIN_00(	4	15			0	0	OKC	13851	Oklahoma	HOU	12191	Texas	419	Southwest	WN	19393	N7858A	
TRAIN_00(	8	15	740	1024	0	0	ORD	13930	Illinois	SLC	14869	Utah	1250	SkyWest A	UA	20304	N125SY	
TRAIN_00(	9	6	1610	1805	0	0	CLT	11057	North Carc	LGA	12953	New York	544	American .	AA	19805	N103US	
TRAIN_00(	7	10	905	1735	0	0	LAX	12892	California	EWR	11618	New Jerse	2454	United Air	UA		N595UA	
TRAIN_00(	1	11	900	1019	0	0	SFO	14771	California	ACV	10157	California	250	SkyWest A	UA	20304	N161SY	
TRAIN_00(	4	13	1545		0	0	EWR	11618		DCA	11278	Virginia	199	Republic A	UA	20452	N657RW	Not_Delayed
TRAIN_00(	1	20	1742	1903	0	0	EWR	11618	New Jerse	BOS	10721	Massachus	200	United Air	UA		N66825	Not_Delayed
TRAIN_00(	4	20	1815	1955	0	0	ORD	13930	Illinois	MCI	13198	Missouri	403		UA	20304	N110SY	
TRAIN_00(	6	13	1420	1550	0	0	BWI	10821		CLT	11057	North Carc	361	Southwest	WN	19393	N765SW	Not_Delayed
TRAIN_00(	6	6	650	838	0	0	LIT	12992	Arkansas	IAH	12266	Texas	374	ExpressJet	UA	20366	N14902	
TRAIN_00(	8	13	1730	1844	0	0	DCA	11278	Virginia	PIT	14122	Pennsylvar	204	Republic A	AA		N119HQ	Delayed

해당 주제 특징 : 항공교통분야 인사이트 도출 가능,구현이 간단함,항공사별 지연건수도 포아송 분포 분석 확장 가능

3.예측 모형 모델링(회귀분석/의사결정나무)

주제 : 야구 승리 예측

사용 데이터 : 1871년부터 2015년까지 야구 데이터

The History of Baseball 의 Team.csv,Batting.csv,Pitching.csv(3개 파일 합쳐야함)

- Team.csv(야구팀 관련 정보,승리,패배 등의 정보 들어있음)

year	league_id	team_id	franchise_i	div_id	rank	g	ghome	w	l
1871		BS1	BNA		3	31		20	10
1871		CH1	CNA		2	28		19	9
1871		CL1	CFC		8	29		10	19
1871		FW1	KEK		7	19		7	12
1871		NY2	NNA		5	33		16	17
1871		PH1	PNA		1	28		21	7
1871		RC1	ROK		9	25		4	21

- Pitching.csv(투수 관련 정보)

player_id	year	stint	team_id	league_id	w	l
bechtge01	1871	1	PH1		1	2
brainas01	1871	1	WS3		12	15
fergubo01	1871	1	NY2		0	0
fishech01	1871	1	RC1		4	16
fleetfr01	1871	1	NY2		0	1

- Batting.csv(타자 관련 정보)

player_id	year	stint	team_id	league_id	g
abercda01	1871	1	TRO		1
addybo01	1871	1	RC1		25
allisar01	1871	1	CL1		29
allisdo01	1871	1	WS3		27
ansonca01	1871	1	RC1		25
armstbo01	1871	1	FW1		12

특징 : 최근 인기가 많아진 소재여서 발표에서 우위 점할 수 있음,3개의 파일 합쳐야 함,야구 통계 분석에서 많이 쓰이는 타율(BA),출루율(OBP),장타율(SLG) 등을 직접 계산해서 회귀모형에 추가하면 더 유의미한 결과 낼 수 있음,전처리 및 구현 과정 복잡함