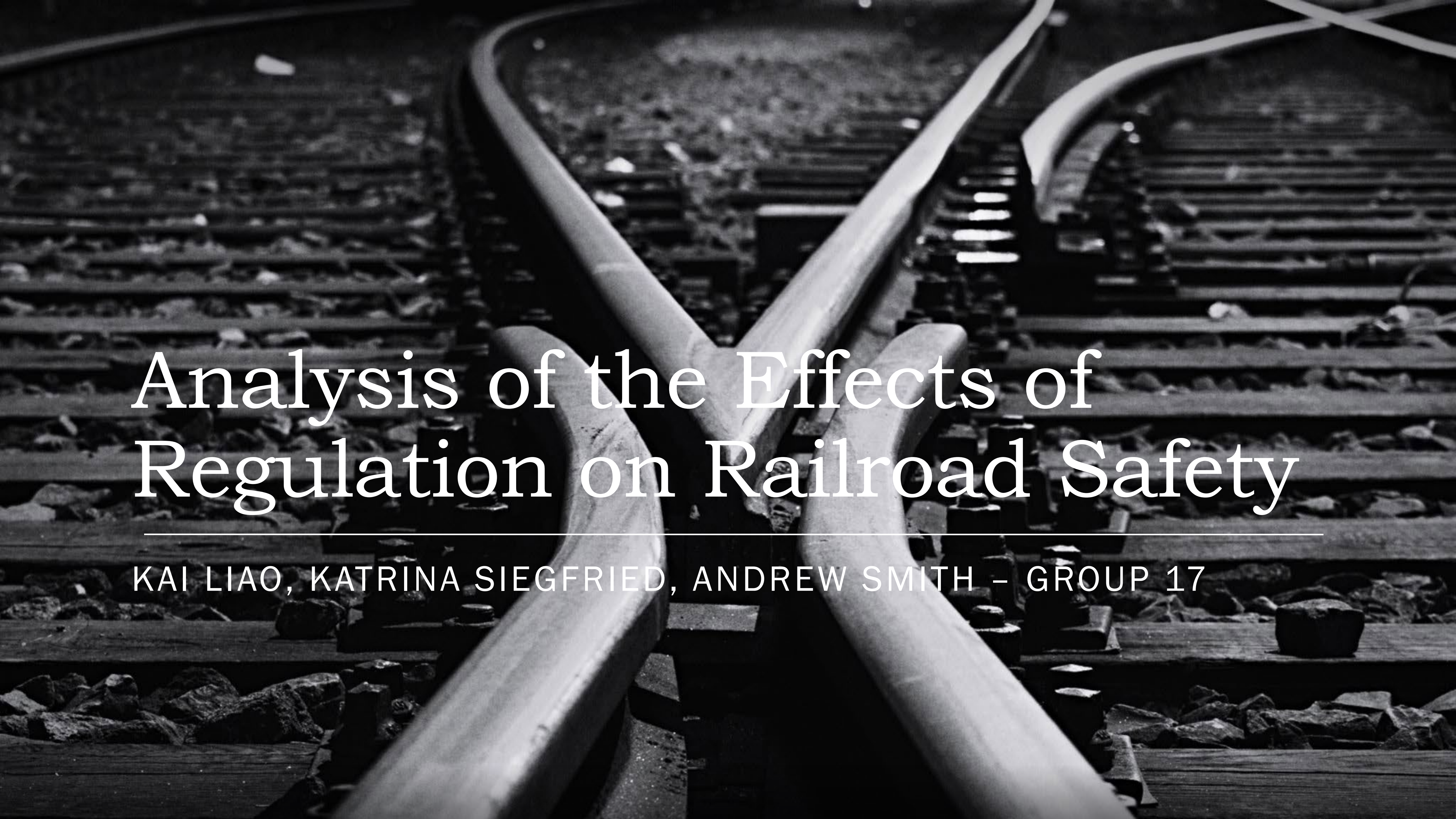

Analysis of the Effects of Regulation on Railroad Safety

KAI LIAO, KATRINA SIEGFRIED, ANDREW SMITH – GROUP 17

Project Description

- Accidents at highway-railroad intersections cause tremendous losses of lives and resources. This project aims to consider the impacts of new regulations, locations of intersections, and the characteristics/topography of intersections to determine which features promote safety and which features do not.
  - Effects of Positive Train Control (PTC) Regulation
  - Accident Analysis by Location
  - Effects of Intersection Characteristics

Project Relevance

June 27, 2022: An Amtrak passenger train struck a dump truck in rural Missouri crossing a passive intersection with no crossing bars, lights and bells.

Several trains and locomotives derailed, 150 people were injured, and 4 killed.

Cost of implementing active restraints would have been $400,000.

There are 130,000 passive railroad crossings nationwide.

Prior Work

Paper	Author	Work Performed	Data Set	Tools	Methods
Analysis of Causes of Major Train Derailment and Their Effect on Accident Rates	Liu et al. (2012)	Identify the causes of train accidents and their effect on accidents rates	Transportation railroads FRA U.S gov. data , (2001–2010 dataset)	Weka tool	Chi-square analysis
Text Mining the Contributors to Rail Accidents	Brown et al. (2016)	Discussed severity of road accident survey.	FRA dataset, from 2001 to 2012	Weka tool	text mining techniques, random forests, partial least squares, latent Dirichlet allocation

Bala, Manju & Bhasin, Anshu. (2018). A Review on Analysis of Railway Traffic Accident with Data Mining Techniques. International Journal of Computer Sciences and Engineering. 6. 1251-1256. 10.26438/ijcse/v6i6.12511256.

Prior Work (cont.)

Paper	Author	Work Performed	Data Set	Tools	Methods
Identifying vehicle driver injury severity factors at highway-railway grade crossings using data mining algorithms	Ghomi et al. (2017)	Identify injury of drivers in factors of highway-railway grade crossing.	U.S (FRA) accident dataset the period of 2006–2013	STATA13 software package tool	Chi-square analysis
Analyzing the Train Accident Injuries using Mining Techniques	Nireesa et al. (2017)	Describe techniques to identify characteristics of accident.	11 years 2001 to 2012 rail accidents U.S. Data set	Weka tool	Association rules, Apriori algorithm

Bala, Manju & Bhasin, Anshu. (2018). A Review on Analysis of Railway Traffic Accident with Data Mining Techniques. International Journal of Computer Sciences and Engineering. 6. 1251-1256. 10.26438/ijcse/v6i6.12511256.

Prior Work (cont.)

Paper	Author	Work Performed	Data Set	Tools	Methods
Human and Organizational Factors of Positive Train Control Safety System	Khashe et al. (2019)	Examining PTC implementation success using organizational safety method	U.S (FRA) accident dataset the period of 1996–2015	General statistical methods	High Reliability Organizing (HRO)
A Novel Data Mining Approach for Analysis of Accident Paths and Performance Assessment of Risk Control Systems	Singh et al. (2020)	Analyze accident paths from incident data and assess performance	Occupational accidents in a steel manufacturing plant in India	Python	Temporal- frequent-, elevated-severity-, and high-impact itemset generation

Dataset

US Department of Transportation – Federal Railroad Administration – Office of Railroad Safety

Highway-Rail Crossing Database

- 186 Attributes
- 436,498 Entries
- 42,567,011 non-empty entries

Collection of all Reports from all US Highway-Rail incidents between 1970-May 2022

https://safetydata.fra.dot.gov/OfficeofSafety/publicsite/DownloadCrossingInventoryData.aspx

- All members have the data downloaded

Proposed Work

Data Cleaning:

1. Correct spelling in narratives, make all letters capital.

2. Ensure all attributes don't have duplicate items representing the same thing

3. Address missing values for each attribute

BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY
RrNarr	StNarr	PolCont	RrCont	HwyCont	THRReque	CCMQzid	Operating	Operating	RrDiv	RrSubDiv	Branch	PrfxMileP	MilePost	SfxMilePc
		8E+09		5.03E+09			NS	Primary	KENTUCKY				81.3	
		8E+09	8.01E+09	5.03E+09			NS	Primary	KENTUCKY				81.3	
PRODUCE	PRODUCE	8.01E+09	8.01E+09	5.03E+09			NS	Primary	MIDWEST	LOUISVILL	#N\A		281.97	W
		8.01E+09	8.01E+09	5.03E+09			NS	Primary	LAKE	CNO&TP N	#N\A		69.48	
RAIL CUT E	Appears a	8.01E+09	8.01E+09	5.03E+09			NS	Primary	MIDWEST	LOUISVILL	IND LEAD		282.5	W
		8.01E+09	8.01E+09	5.03E+09			NS	Primary	MIDWEST	CNO&TP N	#N\A		16.75	
	CROSSING	8.01E+09	8.01E+09	5.03E+09			NS	Primary	GULF	KNOXVILL	CLEAR FORK BR		83.12	C
		8.01E+09	8.01E+09	5.03E+09			NS	Primary	GULF	KNOXVILL	CLEAR FORK BR		83.2	C
		8.01E+09	8.01E+09	5.03E+09			NS	Primary	GULF	KNOXVILL	CLEAR FORK BR		83.5	C
CROSSING	CROSSING	8.01E+09	8.01E+09	5.03E+09			NS	Primary	PIEDMON	KNOXVILL	CLEAR FORK BR		83.7	C
CROSSING	CROSSING SHOULD B		4.05E+09	5.03E+09			NS	Primary	TENNESSE	KNOXVILL	CLEAR FORK BR.		84.28	
		8E+09		5.03E+09			NS	Primary	TENNESSE	KNOXVILL	CLEAR FO	C	84.7	
CROSSING	CROSSING	8E+09	4.05E+09	5.03E+09			NS	Primary	TENNESSE	KNOXVILL	CLEAR FORK BR.		84.5	
		8E+09		5.03E+09			NS	Primary	TENNESSE	KNOXVILL	CLEAR FO	C	84.88	
		8E+09		5.03E+09			NS	Primary	TENNESSE	KNOXVILL	CLEAR FO	C	84.92	
		8E+09		5.03E+09			NS	Primary	TENNESSE	KNOXVILL	CLEAR FO	C	84.98	
APPEARS T	Appears t	8.01E+09	8.6E+09	5.03E+09			RJCC	Primary	CENTRAL I	#N\A	VERSAILLE	LL	11.3	
		8.01E+09	8.01E+09	5.03E+09			NS	Primary	MIDWEST	LOUISVILL	#N\A		289.73	W
		8.01E+09	8.01E+09	5.03E+09			NS	Primary	MIDWEST	LOUISVILL	#N\A		289.9	W
		8.01E+09	8.01E+09	5.03E+09			NS	Primary	MIDWEST	LOUISVILL	#N\A		290.75	W

Proposed Work (cont.)

Data Preprocessing:

PTC implementation:

1. Data reduction using domain "expert" selections.
2. Data transformation by selectively smoothing, normalizing, discretizing data as needed; concept hierarchy generation to mine at different abstraction values; vertical format for FP-growth; one-hot encoding.

Crossing Location:

1. Fill in blanks with identical placeholders or inferred data (if possible).
2. Verify identical formatting for all nominal values and choose relevant orders for ordinal values.

Crossing Characteristics:

1. Specify crossing characteristic attributes as necessary attributes to compare to.
2. Remove irrelevant administrative and reserved attributes.

Tools to Use

Tools:

- Python toolboxes
    - pandas: data cleaning
    - numpy: data transformation
    - sklearn: fix encoding if needed (one-hot), test-train split, random forest
    - pyspark: fp-growth

Methods:

- Decision Trees
- FP-Growth

Evaluation

Method	Details	Evaluation
Decision Tree	Random Forest RI with bagging, will also help with attribute selection, use to generate if-then rule based model	80-20 train/test split, sampling without replacement; accuracy, sensitivity, precision, specificity, F1, Fb
FP-Growth	Explore vertical format to expedite itemset generation; rule pruning	Support, lift, confidence, X2, Kulczynski measure, cosine; final selection of measure dependent on performance on data