OliverWyman

DIY RISK MANAGEMENT DATA EXPLORATION

Bryce Chamberlain, Sr. Manager, Oliver Wyman Actuarial Consulting

A business of Marsh McLennan

RIMS2022 RISKWORLD ANNUAL CONFERENCE & EXHIBITION APRIL 10–13 | SAN FRANCISCO

LEARNING OBJECTIVES

At the end of this session, you will:

1. **Be familiar with tools commonly used for data exploration.**
2. **Understand features and differentiators to determine when each tool should be used.**
3. **Know how to streamline your data exploration and get more value from data.**

DATA WHISPERER, MUSIC ENTHUSIAST

- Associate of the Society of **Actuaries**
- Master of Science in **Analytics**, University of Chicago
- **12 years** professional experience
- Lead team at **Oliver Wyman Actuarial Consulting** building business intelligence apps for the web using **R Shiny**
- Passions: data **visualization**, user-friendly **design**, **efficiency** and **flexibility**

RIMS2022 RISKWORLD
ANNUAL CONFERENCE & EXHIBITION
APRIL 10–13 | SAN FRANCISCO

TOOLS OF THE TRADE

TOOLS OF THE TRADE

Worksheets
Calculator
Excel, Google Sheets

Most people can use it. Easily manipulate single records.	Easy to make mistakes. Difficult to automate. Slow on large data.

Business Intelligence
Get started now
Power BI, Tableau, Salesforce

Lots of options quickly. Click & drag	Limited functionality. Difficult to automate.

AutoML
Search for insights
Rapidminer, DataRobot, SageMaker

Find stories across all data.	Limited visualization. Results are complex. Expensive if not open source.

Design
Make it pretty
Adobe Illustrator, Inkscape

Lots of features and options for perfecting the visual.	Very time consuming. Software is complex, difficult to learn.

Free-hand Drawing
Begin with the end in mind
Pen & Paper, Tablet

Fastest method, no interface to slow you down.	Not generated by data. Not fit for delivery.

Code-based
Automate repetitive tasks
RMarkdown, R Shiny

Unlimited functionality. Open source. Git version control.	Time consuming. Need to code.

SECRET RECIPE

AutoML
Let the computer do the work for you.

Found your story?

Free-hand Drawing
Quickly prototype your final product.

https://www.ft.com/vocabulary

Know what you need?

Business Intelligence
Prototype with data.

Extra mile, one time?

Design

Need to automate?

Code-based

Doesn't fit

Worksheets

CASE STUDY

Our Data

- Auto claims dataset
- https://www.kaggle.com/mykeysid10/insurance-claims-fraud-detection
- 1,000 claims, 39 columns

GLIMPSE YOUR DATA

https://bit.ly/rims-dict

easyr::dict()

Filter

	Name	Class	# Distinct	Top Values	Bottom Values	# Missing	% Missing	% =0	# <0	% <0	Mode	Average
10	insured_sex	character	2	FEMALE, MALE	FEMALE, MALE	0	0	NA	NA	NA	NA	
37	fraud_reported	character	2	N, Y	N, Y	0	0	NA	NA	NA	NA	
4	policy_state	character	3	IL, IN, OH	IL, IN, OH							
6	policy_deductable	integer	3	500, 1000, 2000	500, 1000, 2000	0	0	0.0000	0	0.0000	500.00	
26	property_damage	character	3	?, NO, YES	?, NO, YES	0	0	NA	NA	NA	NA	
27	bodily_injuries	integer	3	0, 1, 2	0, 1, 2	0	0	0.3393	0	0.0000	0.00	
29	police_report_available	character	3	?, NO, YES	?, NO, YES	0	0	NA	NA	NA	NA	
42	incident_date_month	integer	3	1, 2, 3	1, 2, 3							
18	collision_type	character	4	?, Front Collision, Rear Collision, Side Collision	?, Front Collision, Rear Collision, Side Collision	0	0	NA	NA	NA	NA	
19	incident_severity	character	4	Major Damage, Minor Damage, Total Loss, Trivial...	Major Damage, Minor Damage, Total Loss, Trivial...	0	0	NA	NA	NA	NA	
25	number_of_vehicles_involved	integer	4	1, 2, 3, 4	1, 2, 3, 4	0	0	0.0000	0	0.0000	1.00	
28	witnesses	integer	4	0, 1, 2, 3	0, 1, 2, 3	0	0	0.2492	0	0.0000	0.00	
21	incident_state	character	7	NC, NY, OH, PA, SC	OH, PA, SC, VA, WV							
22	incident_city	character	7	Arlington, Columbus, Hillsdale, Northbend, Nort...	Hillsdale, Northbend, Northbrook, Riverwood, Sp...							
22	incident_city	character	7	Arlington, Columbus, Hillsdale, Northbend, Nort...	Hillsdale, Northbend, Northbrook, Riverwood, Sp...	0	0	NA	NA	NA	NA	
41	policy_bind_date_dayofweek	integer	7	1, 2, 3, 4, 5	3, 4, 5, 6, 7	0	0	0.0000	0	0.0000	1.00	
44	incident_date_dayofweek	integer	7	1, 2, 3, 4, 5	3, 4, 5, 6, 7	0	0	0.0000	0	0.0000	1.00	
8	umbrella_limit	integer	11	-1000000, 0, 2000000, 3000000, 4000000	6000000, 7000000, 8000000, 9000000, 10000000	0	0	0.7948	1	0.0010	-1000000.00	111
39	policy_bind_date_month	integer	12	1, 2, 3, 4, 5	8, 9, 10, 11, 12	0	0	0.0000	0	0.0000	1.00	

AUTOML

https://bit.ly/rims-storyteller

```
dt %<>%
    clean(run_autotype = FALSE) %>%
    dropoutliers() %>%
    convert_date_features() %>%
    dropnoisecols() %>%
    correlatedfeatures_find()
```

```
[1] "Checking for outliers."
        dropped [10] rows with outlier at [policy_annual_premium] > 1852.745
        dropped [4] rows with outlier at [capital-gains] > 91516
        dropped [4] rows with outlier at [capital-loss] > 90918
        dropped [8] rows with outlier at [total_claim_amount] > 104064
        dropped [3] rows with outlier at [injury_claim] > 18695
        dropped [7] rows with outlier at [property_claim] > 20250.3
        dropped [9] rows with outlier at [vehicle_claim] > 71508
[1] "Adding features."
[1] "Removing columns that are not useful."
   Dropped column [incident_date_year] for reason [singleval]: 2015.
Searching for high correlation among 946 feature combinations. May take some time.
3% of feature combinations were correlated.
```

```
dt %>%
  correlatedfeatures_address(
    target = 'fraud_reported'
  ) %>%
  fitmodel() %>%
  summary()
```

```
Column [age] dropped and [months_as_customer] kept.
Column [incident_state] dropped and [insured_hobbies] kept.
Column [incident_type] dropped and [collision_type] kept.
Column [collision_type] dropped and [incident_severity] kept.
Column [authorities_contacted] dropped and [incident_severity] kept.
Column [property_damage] dropped and [incident_severity] kept.
Column [injury_claim] dropped and [total_claim_amount] kept.
Column [property_claim] dropped and [total_claim_amount] kept.
Column [total_claim_amount] dropped and [vehicle_claim] kept.
Column [auto_make] dropped and [auto_model] kept.
```

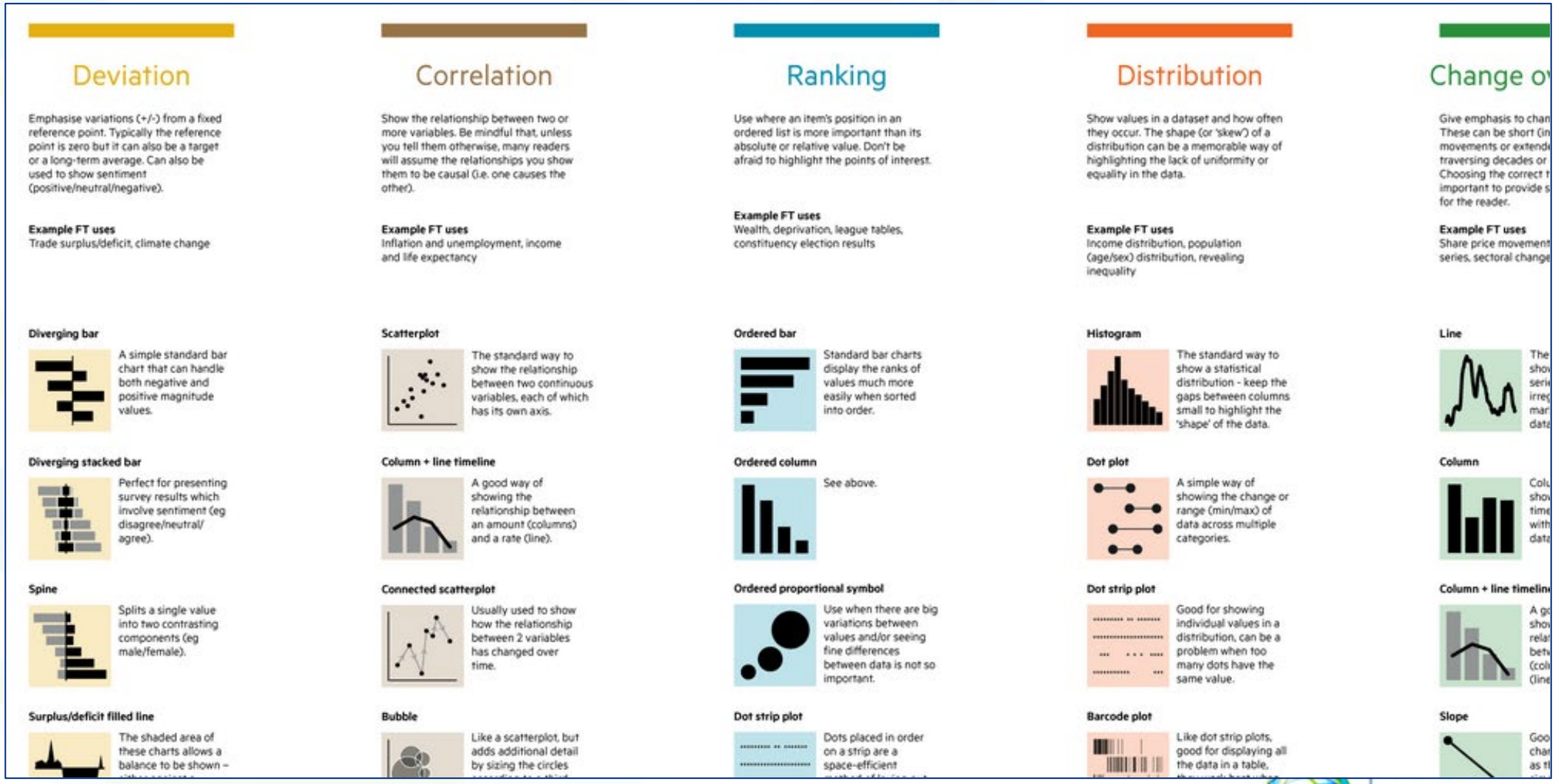
AUTOML

```
dt %>%
  correlatedfeatures_address(
    target = 'fraud_reported'
  ) %>%
  fitmodel() %>%
  summary()
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.164960445036	0.040171694548	4.106	0.0000437	***
umbrella_limit	0.000000010115	0.000000004778	2.117	0.03451	*
policy_bind_date_month	-0.006863287511	0.003146919976	-2.181	0.02943	*
policy_bind_date_dayofweek	-0.008459130645	0.005460581538	-1.549	0.12169	
incident_date_dayofmonth	-0.001376083338	0.001274074058	-1.080	0.28039	
insured_education_level.JD	0.062367733637	0.029905924386	2.085	0.03730	*
insured_occupation.exec.managerial	0.117831435862	0.042060325484	2.801	0.00519	**
insured_hobbies.exercise	-0.086613205931	0.047815216374	-1.811	0.07040	.
insured_hobbies.camping	-0.154213077138	0.048208930769	-3.199	0.00143	**
insured_hobbies.chess	0.608808153878	0.051961628124	11.716	< 0.0000000000000002	***
insured_hobbies.sleeping	-0.155645240007	0.056503966915	-2.755	0.00599	**
insured_relationship.other.relative	0.051752761650	0.029651950080	1.745	0.08125	.
insured_relationship.husband	-0.045365410736	0.029791805908	-1.523	0.12816	
incident_severity.Major.Damage	0.495195916228	0.024774335205	19.988	< 0.0000000000000002	***
auto_model.Wrangler	-0.094068769716	0.054932547630	-1.712	0.08715	.
auto_model.Civic	0.143865911001	0.074785073348	1.924	0.05469	.
auto_model.X6	0.178297750872	0.085898700399	2.076	0.03820	*

FREEHAND DRAWING

https://www.ft.com/vocabulary



Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Example FT uses
Trade surplus/deficit, climate change

Diverging bar
A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar
Perfect for presenting survey results which involve sentiment (eg disagree/neutral/agree).

Spine
Splits a single value into two contrasting components (eg male/female).

Surplus/deficit filled line
The shaded area of these charts allows a balance to be shown –

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other).

Example FT uses
Inflation and unemployment, income and life expectancy

Scatterplot
The standard way to show the relationship between two continuous variables, each of which has its own axis.

Column + line timeline
A good way of showing the relationship between an amount (columns) and a rate (line).

Connected scatterplot
Usually used to show how the relationship between 2 variables has changed over time.

Bubble
Like a scatterplot, but adds additional detail by sizing the circles

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, league tables, constituency election results

Ordered bar
Standard bar charts display the ranks of values much more easily when sorted into order.

Ordered column
See above.

Ordered proportional symbol
Use when there are big variations between values and/or seeing fine differences between data is not so important.

Dot strip plot
Dots placed in order on a strip are a space-efficient

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Example FT uses
Income distribution, population (age/sex) distribution, revealing inequality

Histogram
The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

Dot plot
A simple way of showing the change or range (min/max) of data across multiple categories.

Dot strip plot
Good for showing individual values in a distribution, can be a problem when too many dots have the same value.

Barcode plot
Like dot strip plots, good for displaying all the data in a table,

Change o

Give emphasis to chan
These can be short (in
movements or extend
traversing decades or
Choosing the correct t
important to provide s
for the reader.

Example FT uses
Share price movement
series, sectoral change

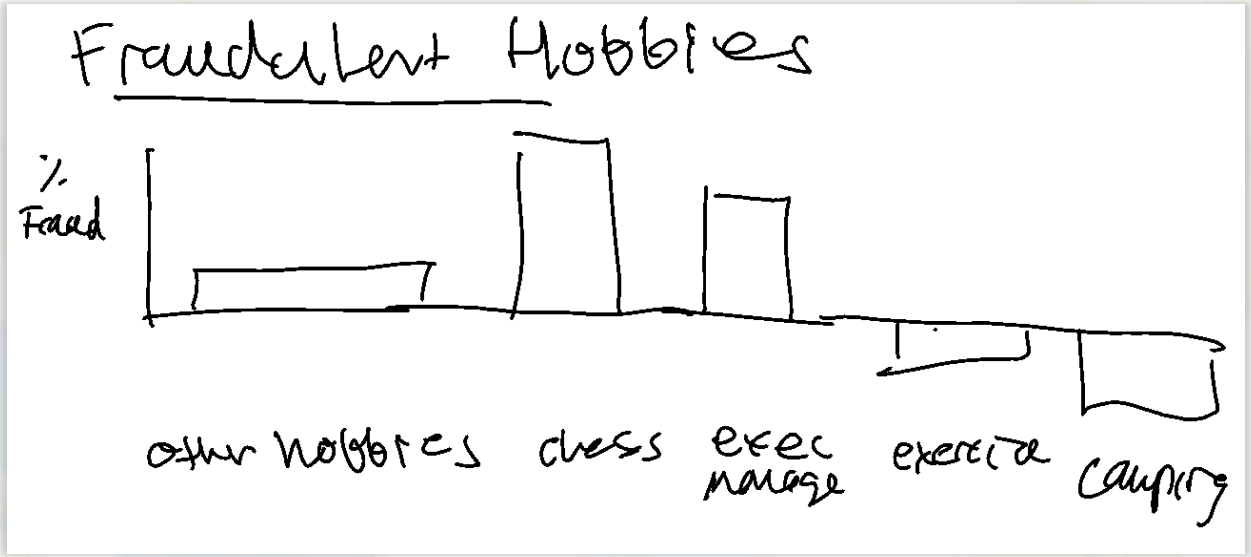
Line
The
shov
serie
irreg
mar
data

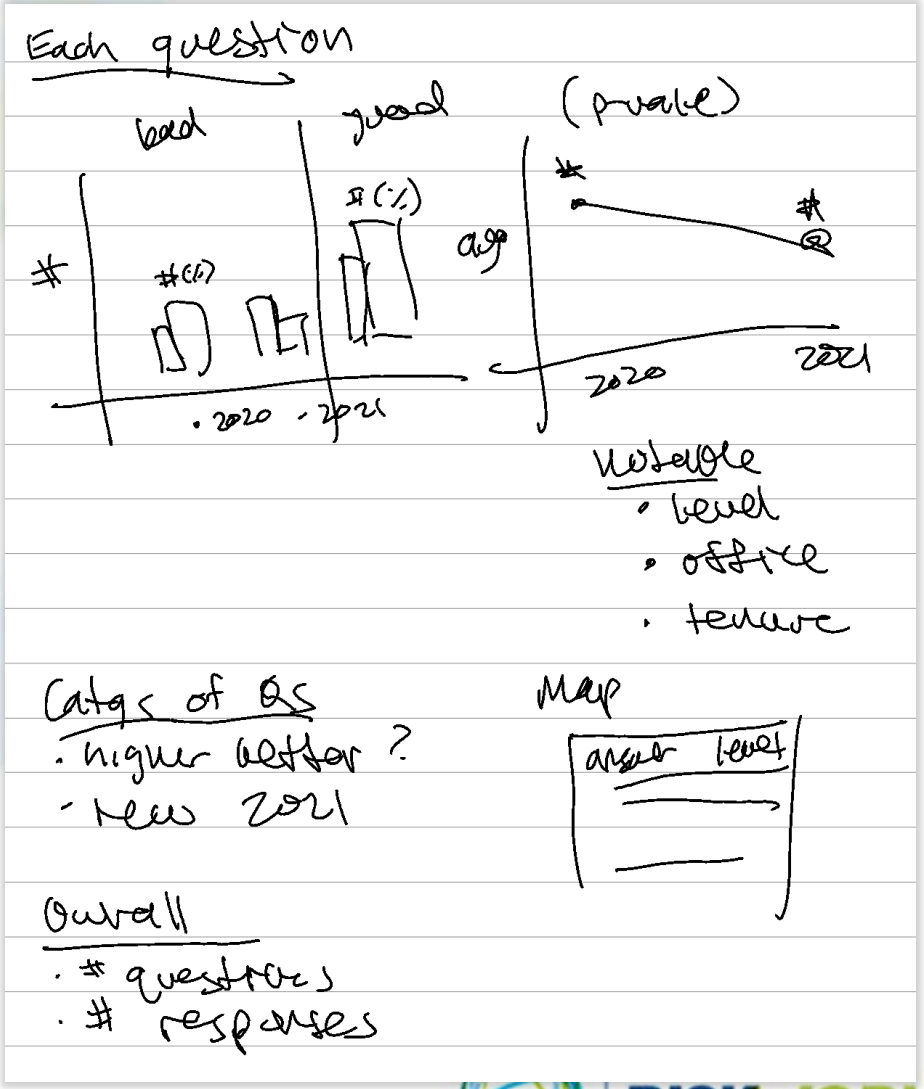
Column
Colu
shov
time
with
data

Column + line timelin
A go
shov
rela
betw
(col
(line

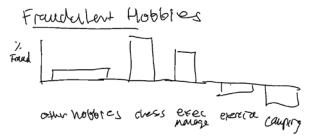
Slope
Goo
char
as t

FREEHAND DRAWING





BUSINESS INTELLIGENCE (BI TOOLS)



pbi - Power BI Desktop

Search

Bryce Chamberlain

File Home Insert Modeling View Help Format Data / Drill Table tools Measure tools

Name: Measure
Home table: model-data
Format: General
Auto
Data category: Uncategorized
New measure
Quick measure

```
1 Measure = AVERAGE('model-data'[fraud1])-0.24
```

Fraudulent Hobbies

% Fraud Reported

Insured's Hobby

chess	cross-fit	board-games	base-jumping	polo	yachting	reading	hiking	skydiving	paintball	video-games	exercise	basketball	movies	bungie-jumping	sleeping	dancing	golf	kayaking	camping
0.58	0.52	0.06	0.03	0.02	0.01	0.00	-0.01	-0.01	-0.03	-0.04	-0.05	-0.06	-0.07	-0.08	-0.08	-0.12	-0.13	-0.14	-0.15

Average of fraud1

0.24

Visualizations

Build visual

Axis: insured_hobbies

Legend: Add data fields here

Values: Measure

Small multiples: Add data fields here

Tooltips: Count of age

Drill through

Cross-report: Off

Keep all filters: On

Add drill-through fields here

Filters

Fields

Search

model-data

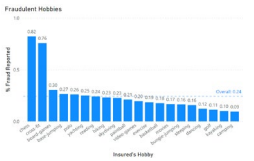
age
authorities_con...
auto_make
auto_model
auto_year
bodily_injuries
capital-gains
capital-loss
collision_type
fraud_reported
fraud1
incident_city
incident_date_...
incident_date_...
incident_date_...
incident_hour_...
incident_location
incident_severity
incident_state
incident_type
injury_claim
insured_educat...
insured_hobbies
insured_occupa...
insured_relatio...
insured_sex
insured_zip
Measure
months_as_cus...

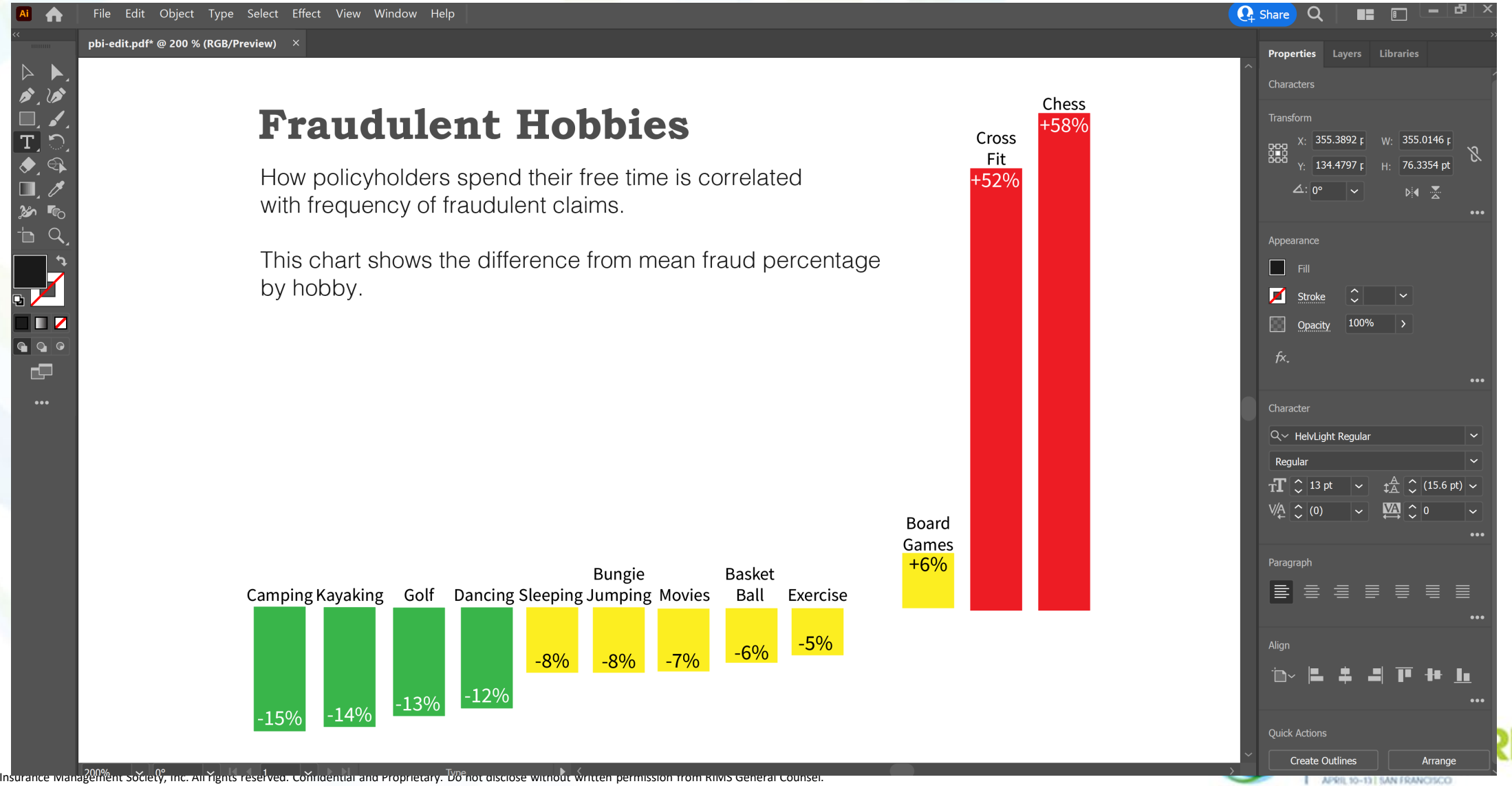
Page 1 Page 2 Page 3 Duplicate of Page 3 Page 4

Page 5 of 5

DESIGN

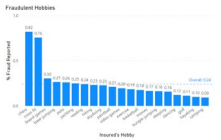
https://bit.ly/rims-illustrator

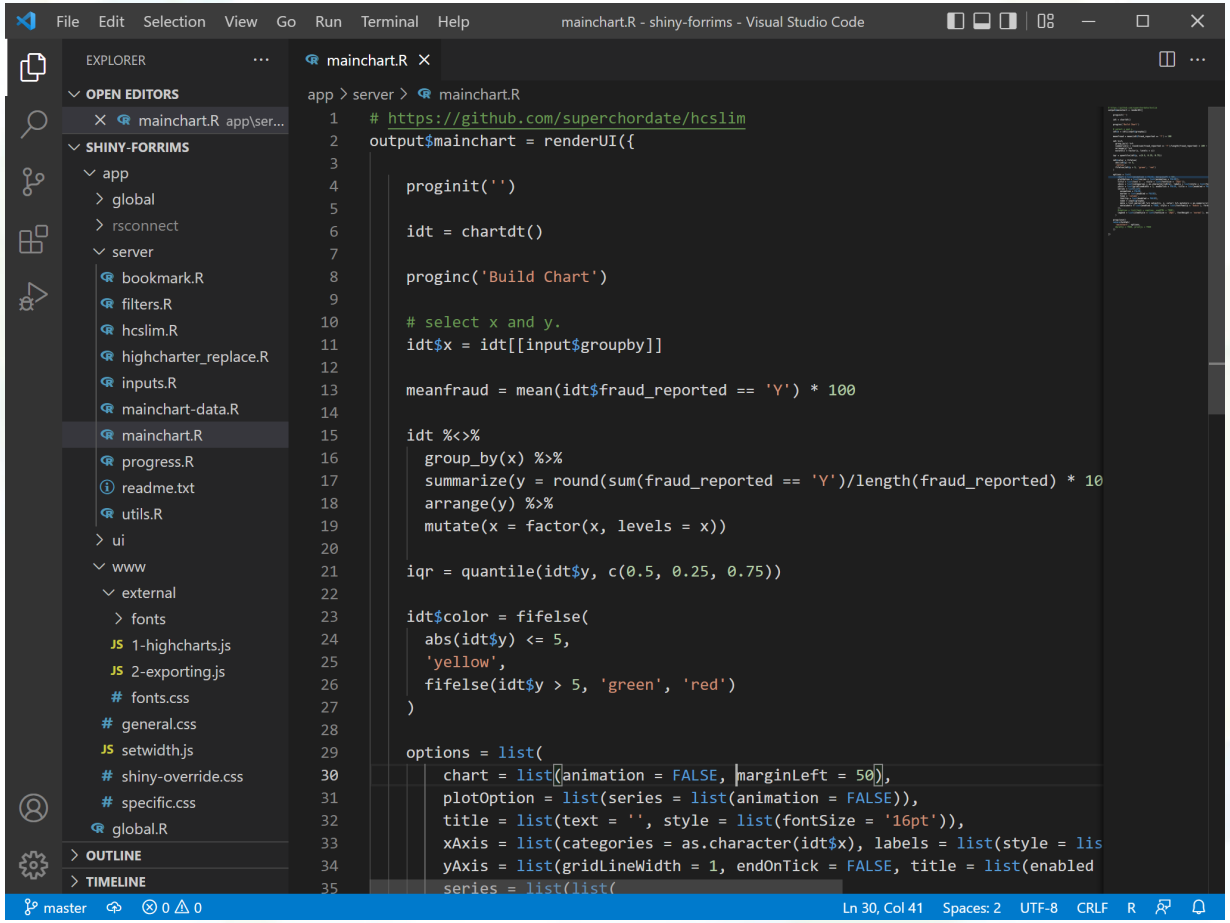


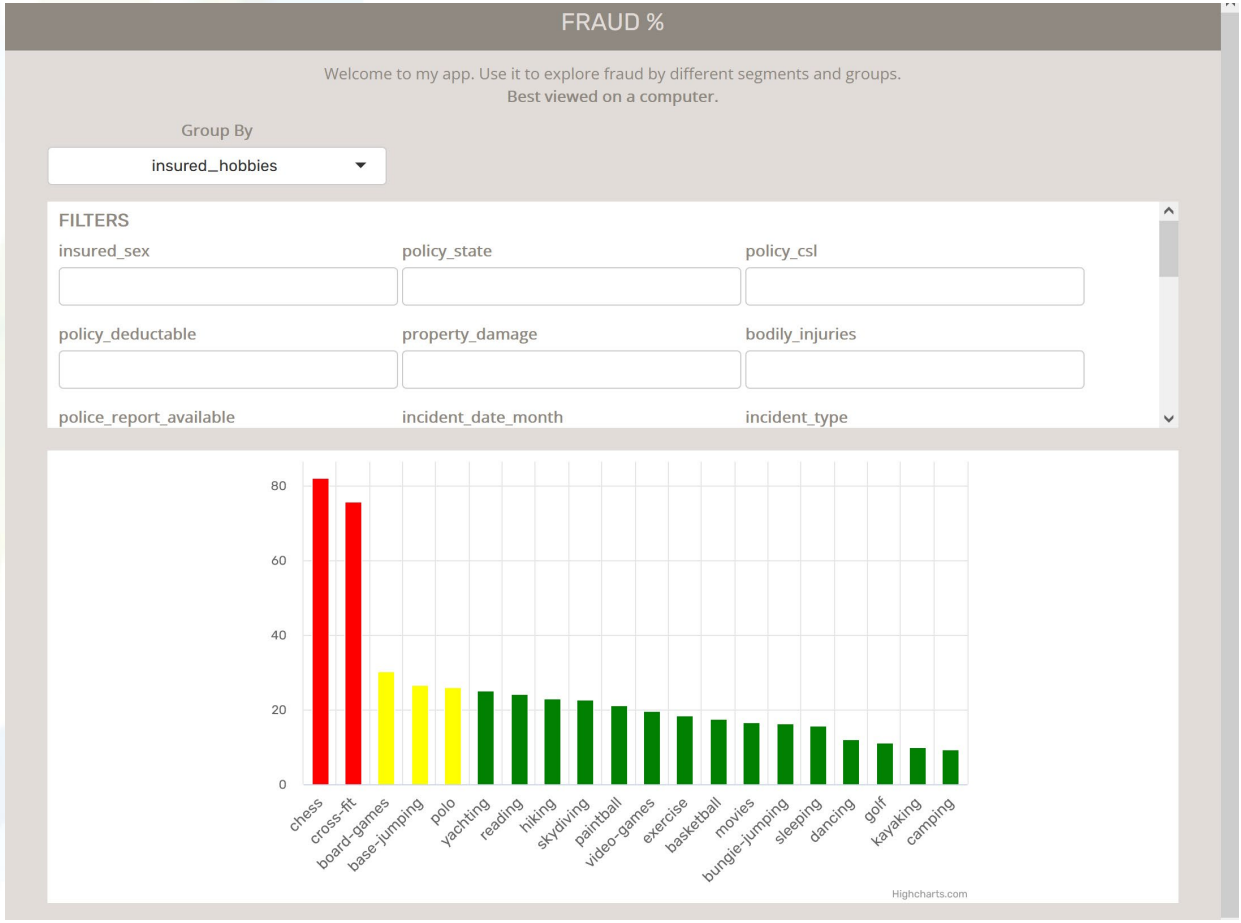


CODEBASED

https://bit.ly/rims-codebased







SECRET RECIPE REFRESHER

AutoML
Let the computer do the work for you.

Found your story?

Free-hand Drawing
Quickly prototype your final product.

https://www.ft.com/vocabulary

Know what you need?

Business Intelligence
Prototype with data.

Extra mile, one time?

Design

Need to automate?

Code-based

REMINDERS, Q&A

Bryce.Chamberlain@oliverwyman.com

linkedin.com/in/brycechamberlain

Hand me your card for a follow-up email including links

Complete Evaluation Rating in the RIMS Mobile App

Your feedback is very important to determine the success and help make improvements

Enjoy your next session!