

AXA x CitiBike



Sayyam Alam

Agenda

1. Einführung
2. Methodisches Vorgehen
3. Ergebnisse
4. Fazit

Partnerschaft AXA x CitiBike

- Nutzer sind bei Unfällen während der Fahrt oft **nicht gesondert abgesichert**.
- CitiBike bietet derzeit **keinen integrierten Schutz** an, der unmittelbar mit der Fahrt verknüpft ist.
- AXA kann einen **optionalen, kontextbezogenen Versicherungsschutz** bereitstellen.
- CitiBike liefert die Daten, AXA den Service.

Produktidee: Micro-Insurance pro Fahrt

Eine Zusatzversicherung, die beim Start einer Fahrt angeboten wird.

Wer profitiert davon?

Nutzer

Können bei Bedarf einfachen Schutz dazubuchen.

CitiBike

Kann sein Angebot um freiwillige Absicherung erweitern.

AXA

Stellt den Versicherungsschutz bereit.

Ziel und Scope

- Für das Produkt braucht es eine Möglichkeit, um das **Unfallrisiko** einer Fahrt einzuschätzen.
- Dafür müssen **potenzielle Risikotreiber** identifiziert werden.
- Gezeigt wird ein möglicher Analysepfad – kein fertiges Produkt oder Modell.

Methodisches Vorgehen

Phase 1:

Projekt-Framing

Phase 2:

Explorative Datenanalyse

Phase 3:

Hypothesentests

Phase 4:

Risikomodellierung

Methodisches Vorgehen

Phase 1:

Projekt-Framing

Phase 2:

Explorative Datenanalyse

Phase 3:

Hypothesentests

Phase 4:

Risikomodellierung

Datenbasis

CitiBike-Fahrtdataen

- Monatliche Fahrtdataen (2023)
- Ca. 35.000.000 Zeilen, 13 Spalten
- Enthält: Start/Ende, Station, Koordinaten, Bike-Typ, etc.

NYPD-Unfalldaten

- Offene Verkehrsunfalldaten (2023) mit Schaden > 1000\$
- Ca. 96.000 Zeilen, 29 Spalten
- Enthält: Ort, Zeit, Art des Unfalls, Beteiligte, Schweregrad, etc.

	ride_id	rideable_type	started_at	ended_at	start_station_name	start_station_id	end_station_name	end_station_id	start_lat	start_lng	end_lat	end_lng	member_casual
0	A8518A6C4BE513DE	classic_bike	2023-01-03 23:14:52.325	2023-01-03 23:33:42.737	E 1 St & Bowery	5636.13	Spruce St & Nassau St	5137.10	40.724861	-73.992131	40.711464	-74.005524	casual
1	A3911E4F5B9B5773	electric_bike	2023-01-07 07:57:40.054	2023-01-07 08:01:27.330	E 1 St & Bowery	5636.13	Ave A & E 11 St	5703.13	40.724861	-73.992131	40.728547	-73.981759	casual
2	AE7F74C32AEBF6F2	electric_bike	2023-01-09 18:37:44.830	2023-01-09 18:48:56.233	1 Ave & E 39 St	6303.01	E 14 St & 1 Ave	5779.10	40.747140	-73.971130	40.731393	-73.982867	member
3	6E10997509D2B7F6	electric_bike	2023-01-05 19:06:15.350	2023-01-05 19:08:33.547	E Burnside Ave & Ryer Ave	8397.02	E Burnside Ave & Ryer Ave	8397.02	40.850535	-73.901318	40.850535	-73.901318	casual
4	AA546E74A9330BD4	electric_bike	2023-01-02 20:25:23.300	2023-01-03 10:51:25.164	Clermont Ave & Park Ave	4692.01	Clermont Ave & Park Ave	4692.01	40.695734	-73.971297	40.695734	-73.971297	casual

	CRASH_DATE	CRASH_TIME	BOROUGH	ZIP_CODE	LATITUDE	LONGITUDE	LOCATION	ON STREET NAME	CROSS STREET NAME	OFF STREET NAME	CONTRIBUTING FACTOR VEHICLE 2	CONTRIBUTING FACTOR VEHICLE 3	CONTRIBUTING FACTOR VEHICLE 4	CONTRIBUTING FACTOR VEHICLE 5	COLLISION_ID	VEHICLE TYPE CODE 1	VEHICLE TYPE CODE 2	TYPE
0	01/01/2023	5:30	None	NaN	40.710514	-73.956140	(40.710514, -73.95614)	BROOKLYN QUEENS EXPRESSWAY	None	None	...	Unspecified	None	None	None	4599014	Sedan	None
1	01/01/2023	8:45	BRONX	10457.0	40.845870	-73.890730	(40.84587, -73.89073)	None	None	1972 CROTONA AVENUE	...	Unspecified	Unspecified	None	None	4598137	Sedan	Sedan
2	01/01/2023	19:00	BROOKLYN	11206.0	40.708237	-73.943370	(40.708237, -73.94337)	None	None	179 GRAHAM AVENUE	...	None	None	None	4599015	Station Wagon/Sport Utility Vehicle	None	
3	01/01/2023	16:35	BROOKLYN	11221.0	40.693660	-73.931540	(40.69366, -73.93154)	None	None	1073 DE KALB AVENUE	...	Unspecified	Unspecified	Unspecified	None	4599499	Station Wagon/Sport Utility Vehicle	Station Wagon/Sport Utility Vehicle

Rahmen der Analyse

Problem

- Nutzung (CitiBike) und Unfallereignisse (NYPD) werden getrennt erfasst und sind nicht direkt verknüpfbar.

Ansatz

- Gemeinsame Vergleichsachsen festlegen.

Raum

- Präzise Koordinaten in beiden Quellen
- Ermöglicht räumliche Gegenüberstellung

Zeit

- Vollständige Zeitstempel in beiden Quellen
- Ermöglicht Vergleich nach Tageszeit / Wochentag / Saison

Methodisches Vorgehen

Phase 1:
Projekt-Framing

Phase 2:
Explorative Datenanalyse

Phase 3:
Hypothesentests

Phase 4:
Risikomodellierung

Struktur von Phase 2

Phase 2A:

Aufbereitung der
CitiBike-Daten

Phase 2A:

Aufbereitung der NYPD-
Daten

Phase 2B:

Räumliche Analyse

Phase 2B:

Zeitliche Analyse



Struktur von Phase 2

Phase 2A:

Aufbereitung der
CitiBike-Daten

Phase 2A:

Aufbereitung der NYPD-
Daten

Phase 2B:

Räumliche Analyse

Phase 2B:

Zeitliche Analyse



CitiBike - Bereinigungsschritte

Strukturprüfung

- Prüfen der Geodaten auf Plausibilität
- Prüfen auf Duplikate und fehlende Werte

Berechnete Variablen

- Berechnung der Fahrtzeit
- Prüfen der Fahrtzeit auf unplausible Werte (≤ 0 Minuten, > 24 Stunden, Endzeit < Startzeit)

Datenreduktion und Konsistenz

- Einschränkung auf Fahrten aus 2023 (zeitliche Konsistenz zu NYPD)
- Entfernen nicht benötigter Attribute (z.B. Stationnamen)

CitiBike - Datenqualität

Entfernte Einträge

- **Unplausible Fahrtdauern**
 - 324 Einträge mit Dauer \leq 0 Minuten
 - 26.640 Fahrten mit Dauer $>$ 24 Stunden
- **Fehlende oder unbrauchbare Koordinaten**
 - 79.644 Einträge

Ergebnis der Bereinigung

- Rohdatenmenge: ca. 35,1 Mio Fahrten
- Verbleibend: ca. 35,0 Mio Fahrten
- Anteil behaltener Daten: **99,7%**

CitiBike - Erste Erkenntnisse

Nutzungstypen

- Member: ca. 28,5 Mio
- Casual: ca. 6,5 Mio

Fahrtdauer

- Median: ca. 9 Minuten

Fahrradtypen

- E-Bike: ca. 17,6 Mio
- Fahrrad: ca. 17,5 Mio

Stationen

- > 2.300 Startstationen
- > 2.300 Endstationen

Struktur von Phase 2

Phase 2A:

Aufbereitung der
CitiBike-Daten

Phase 2A:

Aufbereitung der NYPD-
Daten

Phase 2B:

Räumliche Analyse

Phase 2B:

Zeitliche Analyse



NYPD - Bereinigungsschritte

Strukturprüfung

- Prüfung der Format- und Typkonsistenz (Bereinigung von Whitespaces, einheitliche Schreibweise)
- Prüfen auf Duplikate und fehlende Werte
- Prüfen der Geodaten auf Plausibilität

Berechnete Variablen

- Extraktion der Fahrzeugtypen aus fünf Vehicle-Code-Spalten
- Konstruktion eines Flags *cyclist_involved*
 - Basierend auf Verletzungsdaten und Fahrzeugtypen

NYPD - Datenqualität

Entfernte Einträge

- **Ungültige Geokoordinaten**
 - 7.593 Einträge mit fehlenden oder ungültigen Koordinaten
 - Davon 408 Fälle mit Fahrrad/E-Bike Beteiligung

Ergebnis der Bereinigung

- Rohdatenmenge: 96.606
- Verbleibend: 89.013
- Anteil behaltender Daten: **92,1%**

Charakteristik des bereinigten Datensatzes

- Unfälle mit Fahrrad/E-Bike-Beteiligung: 7.963 (ca. 8,2%)

Struktur von Phase 2

Phase 2A:

Aufbereitung der
CitiBike-Daten

Phase 2A:

Aufbereitung der NYPD-
Daten

Phase 2B:

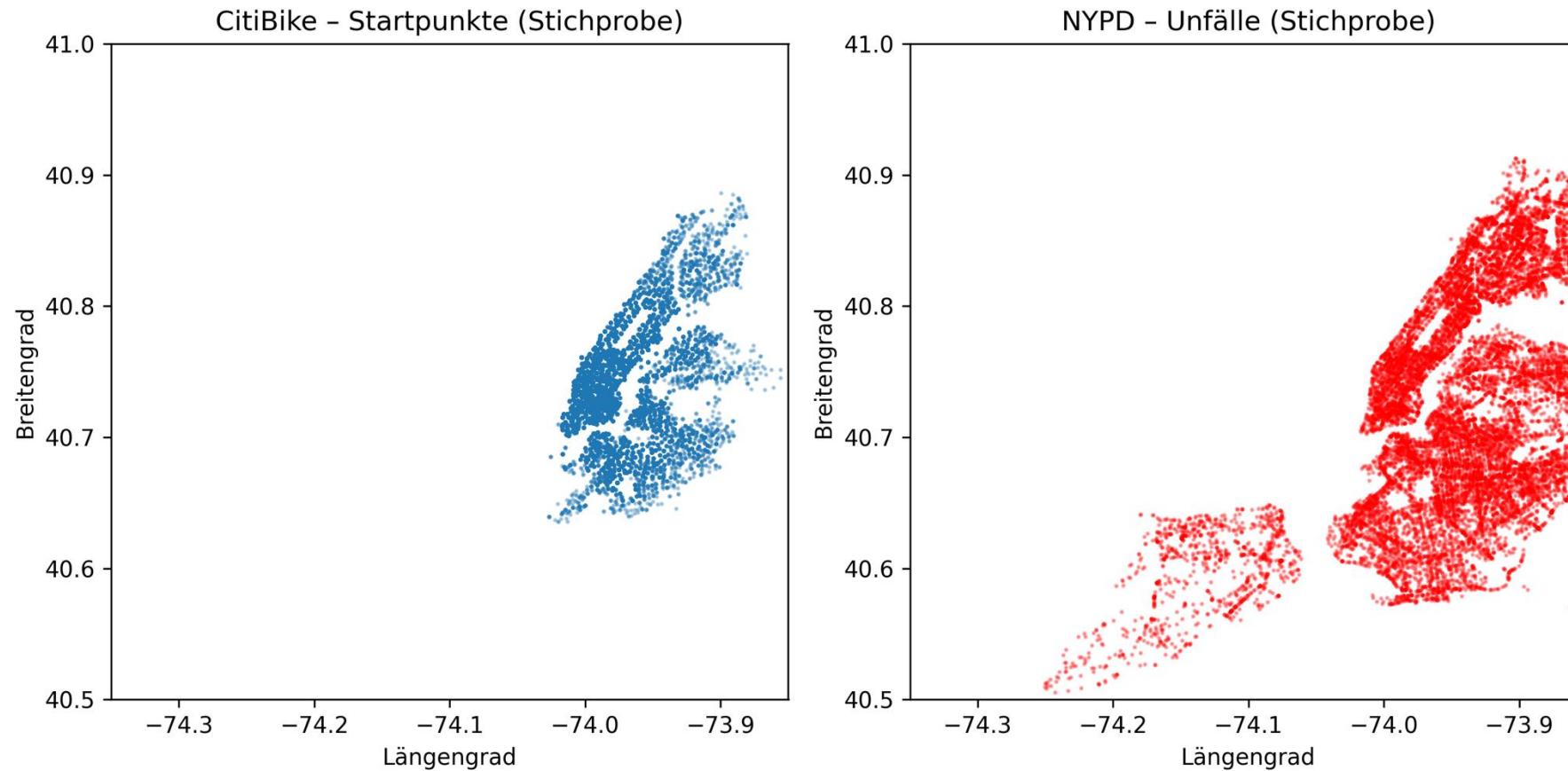
Räumliche Analyse

Phase 2B:

Zeitliche Analyse



Räumliche Diskrepanz zwischen CitiBike und NYPD



Erzeugung eines räumlichen Rasters

- Es wird ein Raster über dem Gebiet erstellt, in dem CitiBike tatsächlich benutzt wird

Warum ein CitiBike-basiertes Raster?

- CitiBike definiert den Exposure-Raum – den Raum, wo Fahrten stattfinden
- NYPD deckt deutlich größere Bereiche ab

Wie sieht das Raster aus?

- Zellgröße 0,003 LAT x 0,003 LON (ca. 300 m)
- Jede Zelle enthält einen eindeutigen Index
- Fahrten und Unfälle werden diesen Zellen zugeordnet

Ergebnis

- 7.134 Zellen mit 100% der CitiBike Fahrten, ca. 69% der NYPD-Unfälle

Räumliche Exposure pro Rasterzelle

Was bedeutet Exposure?

- Exposure misst, wie häufig eine Rasterzelle durch CitiBike-Fahrten „berührt“ wird

Wie berechnet sich Exposure?

- Eine Fahrt kann maximal zwei Zellen berühren: Start und Ende
- Wenn Start \neq Ende: beide Zellen +1
- Wenn Start = Ende: Zelle +1 (keine Doppelzählung)
- Keine Rekonstruktion der Route

Was wird berechnet?

- Für jede Zelle
 - exposure_total
 - exposure_member, exposure_casual
 - exposure_classic, exposure_electric

Kennzahlen zur räumlichen Exposure

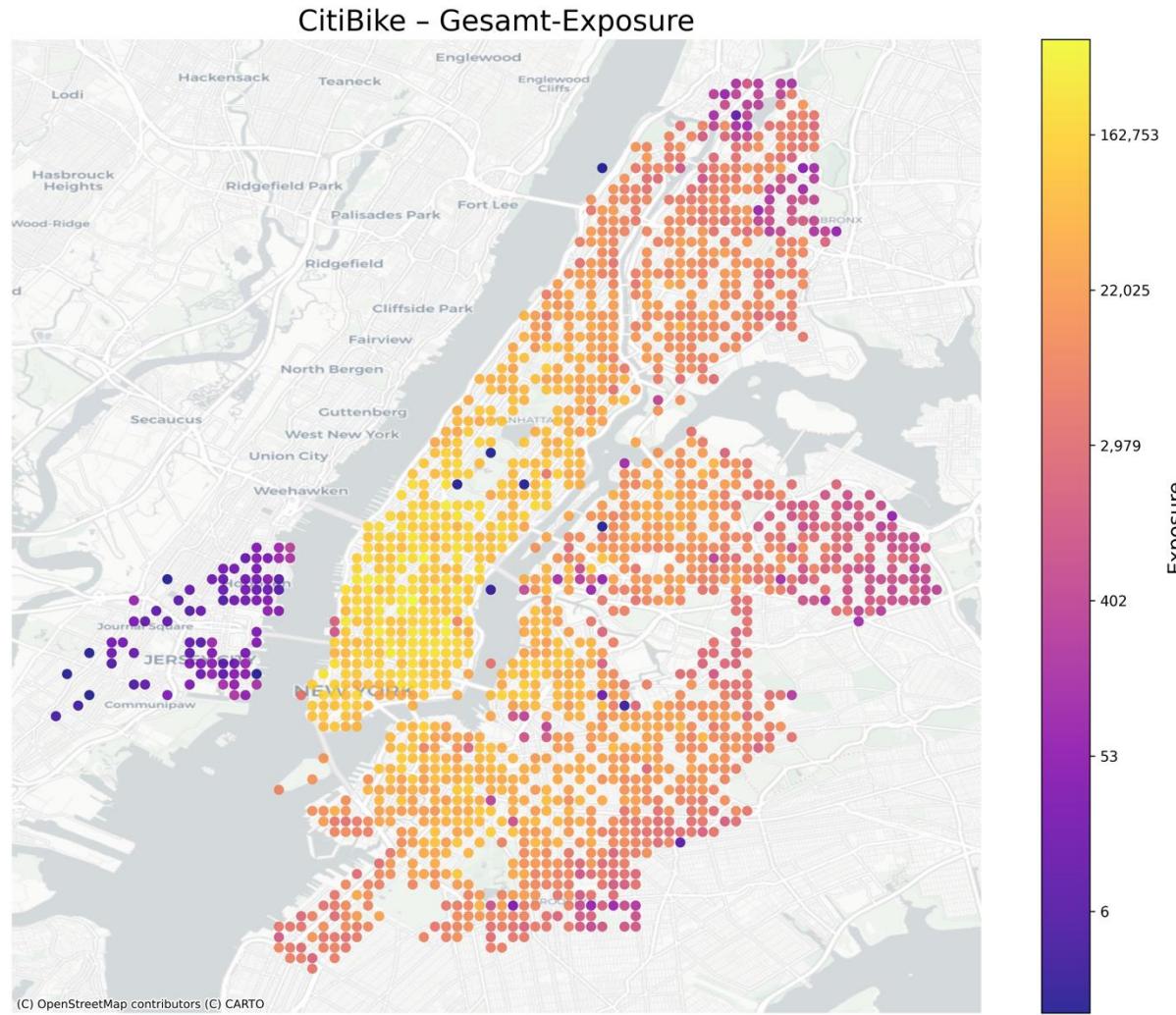
Exposure pro Fahrt

- **3,7 %** der Fahrten starten und enden in derselben Zelle
- $\approx 68,7$ Mio Exposure-Events insgesamt
- $\varnothing 1,96$ Exposure pro Fahrt

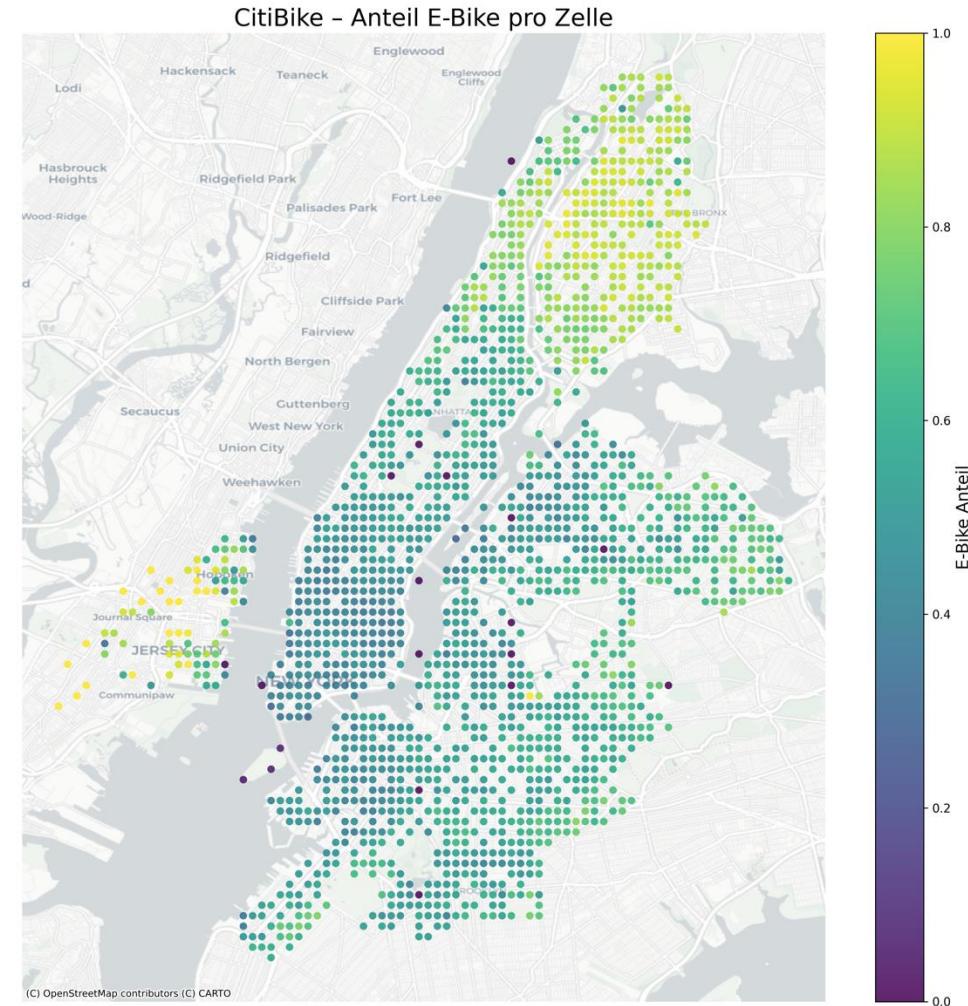
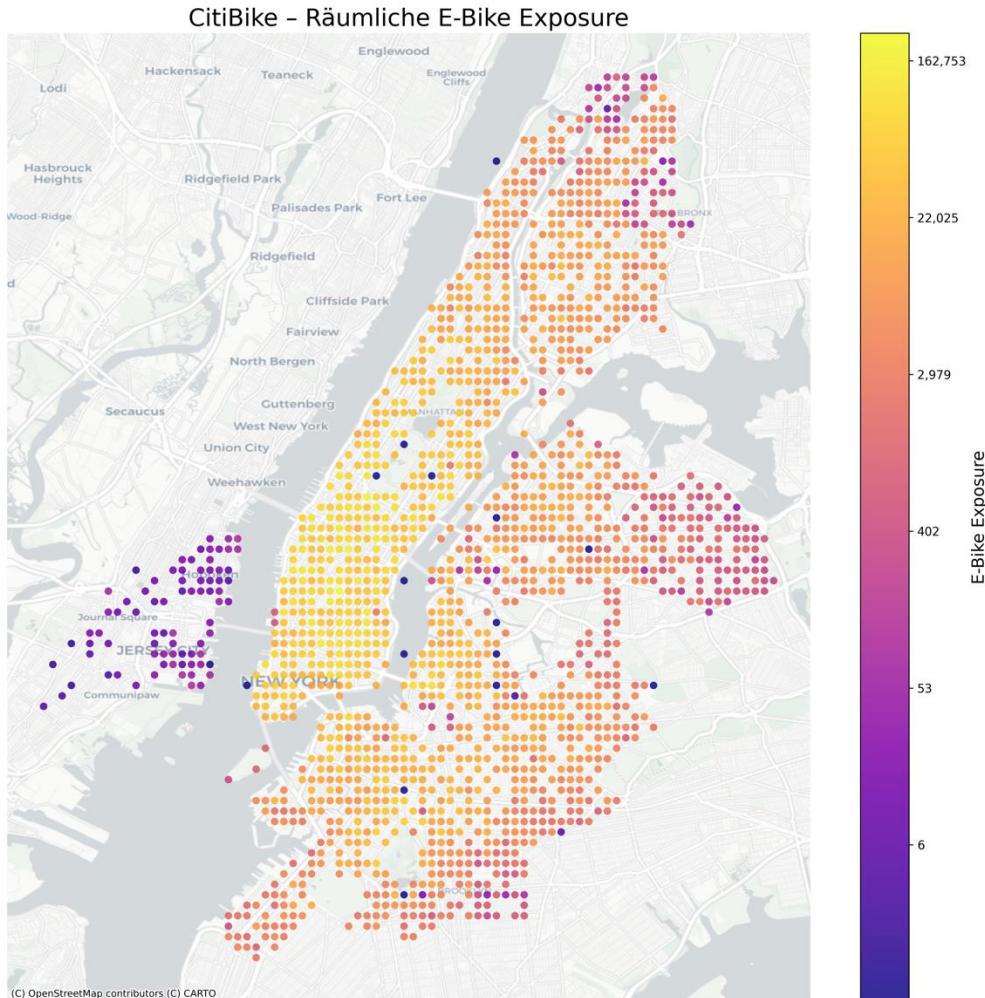
Exposure pro Zelle

- **1.716** Zellen mit positiver Exposure
- Median: ≈ 15.000
- Maximum: ≈ 554.000
- **90%** aller Zellen haben **<114.000** Exposure

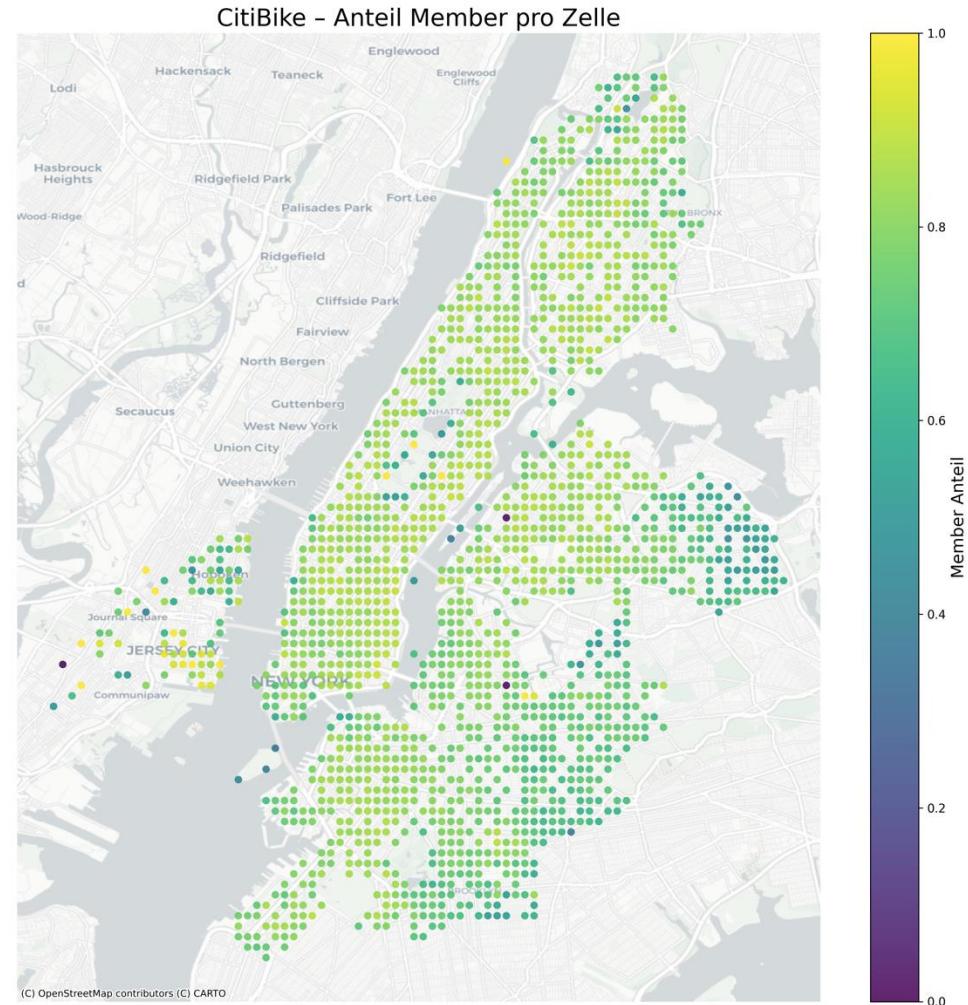
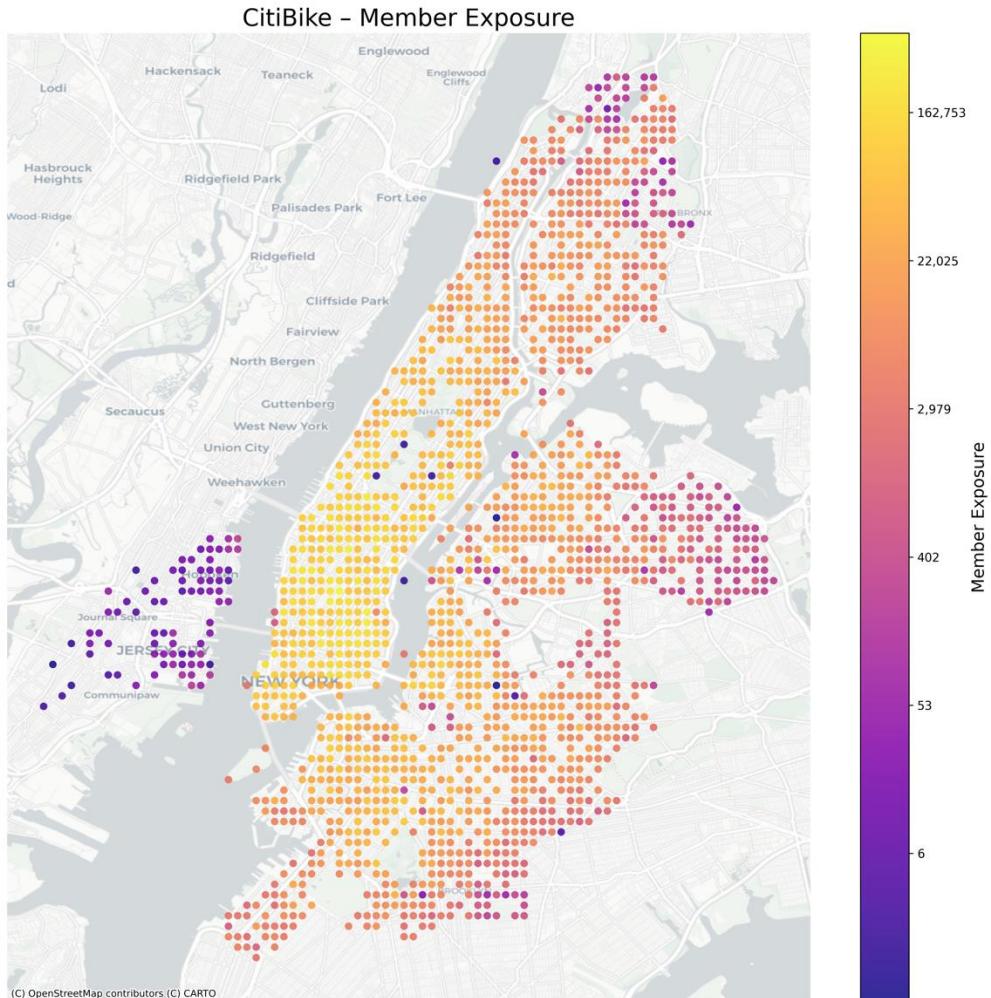
Gesamt-Exposure



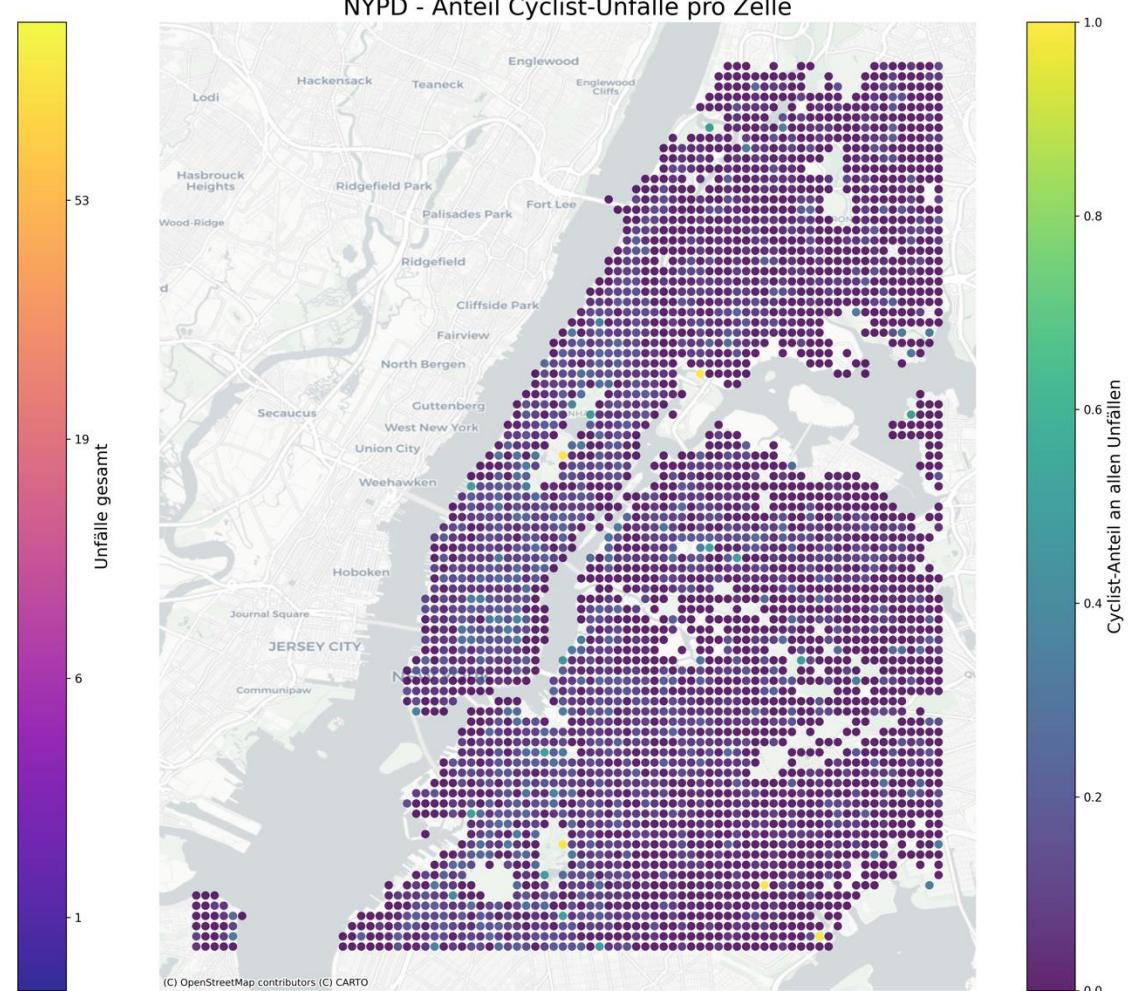
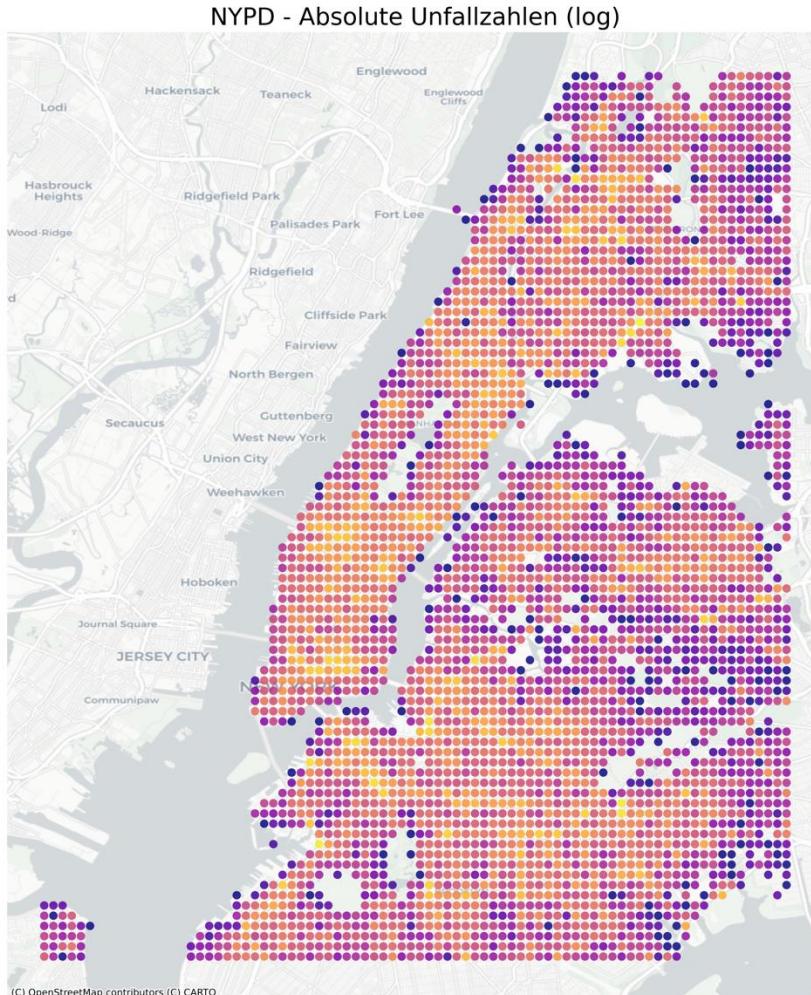
E-Bike Exposure



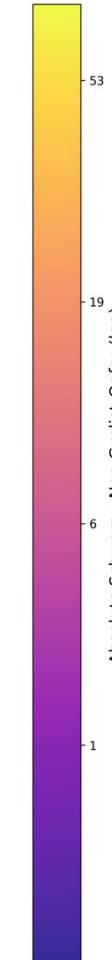
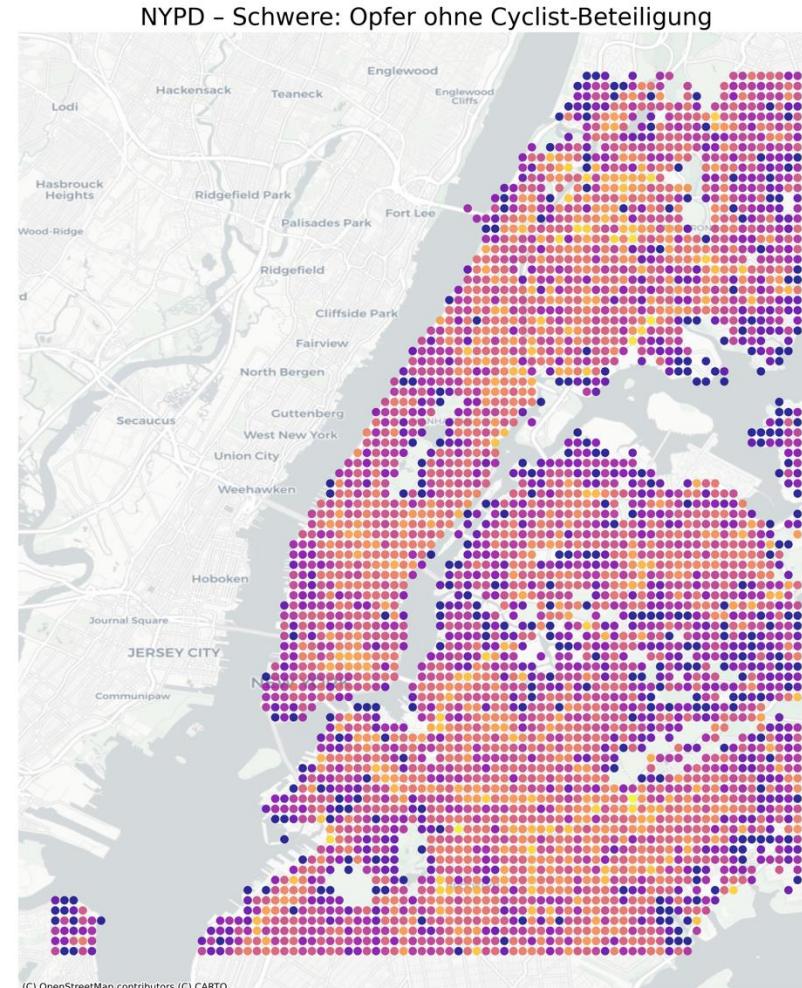
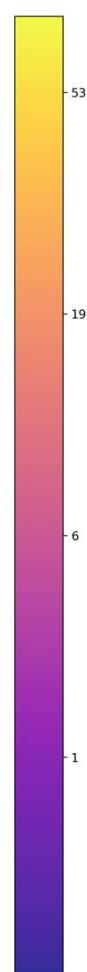
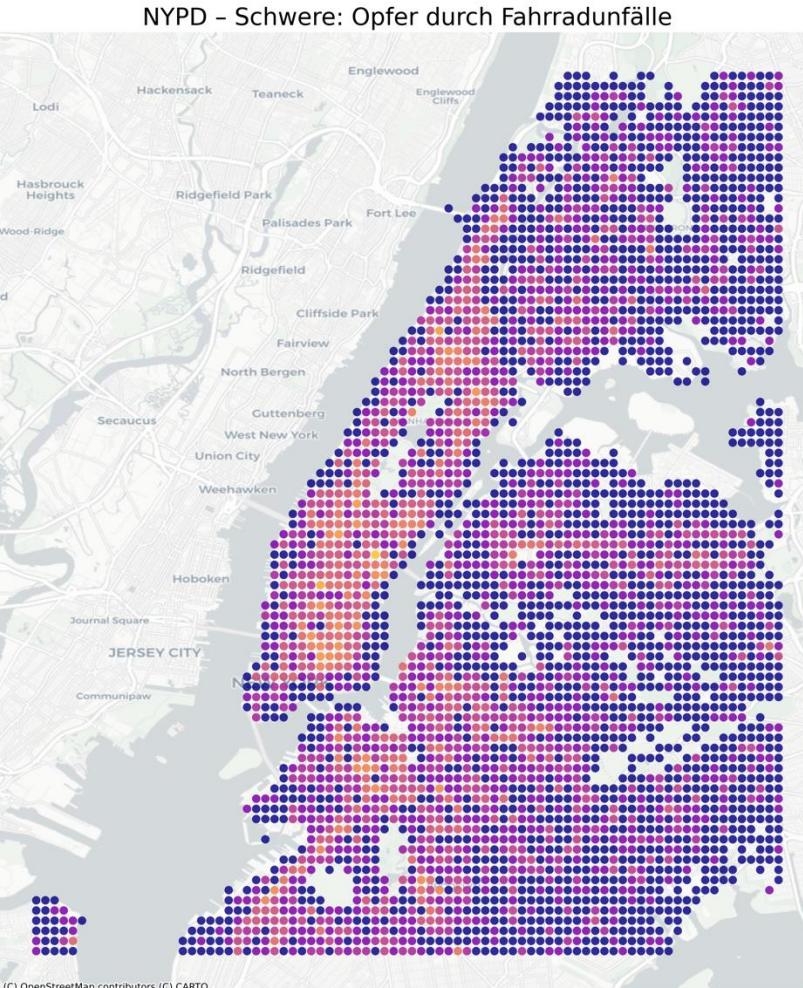
Member Exposure



Absolute Unfallzahlen & Anteil Cyclist-Unfälle

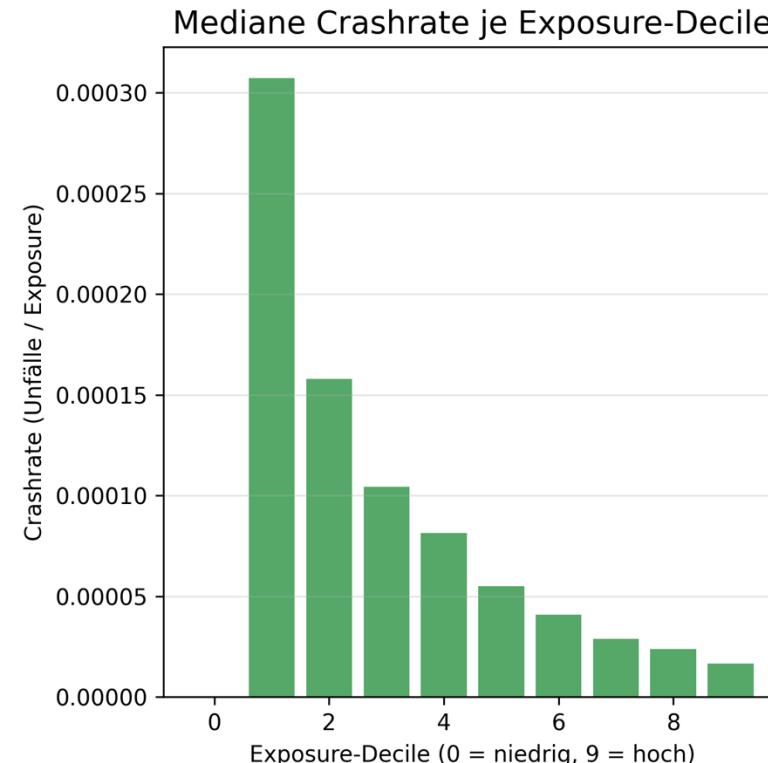
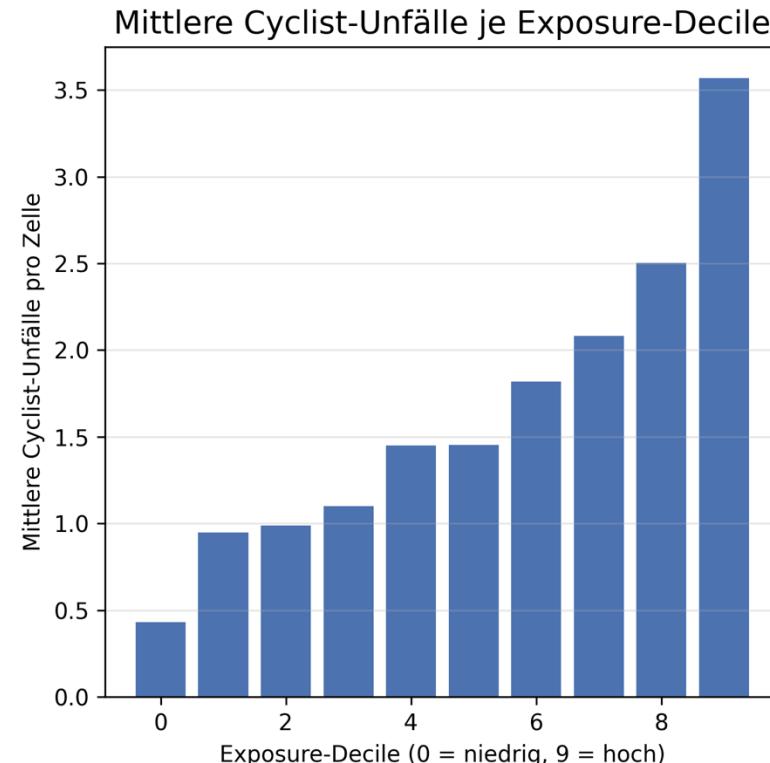


Schwere der Unfälle



Beziehung zwischen Nutzung und Unfallhäufigkeit

- Alle Rasterzellen werden nach ihrer Exposure sortiert
- Anschließend werden sie in 10 gleich große Gruppen eingeteilt – Exposure Deciles
- **Safety-in-Numbers-Effekt**



Struktur von Phase 2

Phase 2A:

Aufbereitung der
CitiBike-Daten

Phase 2A:

Aufbereitung der NYPD-
Daten

Phase 2B:

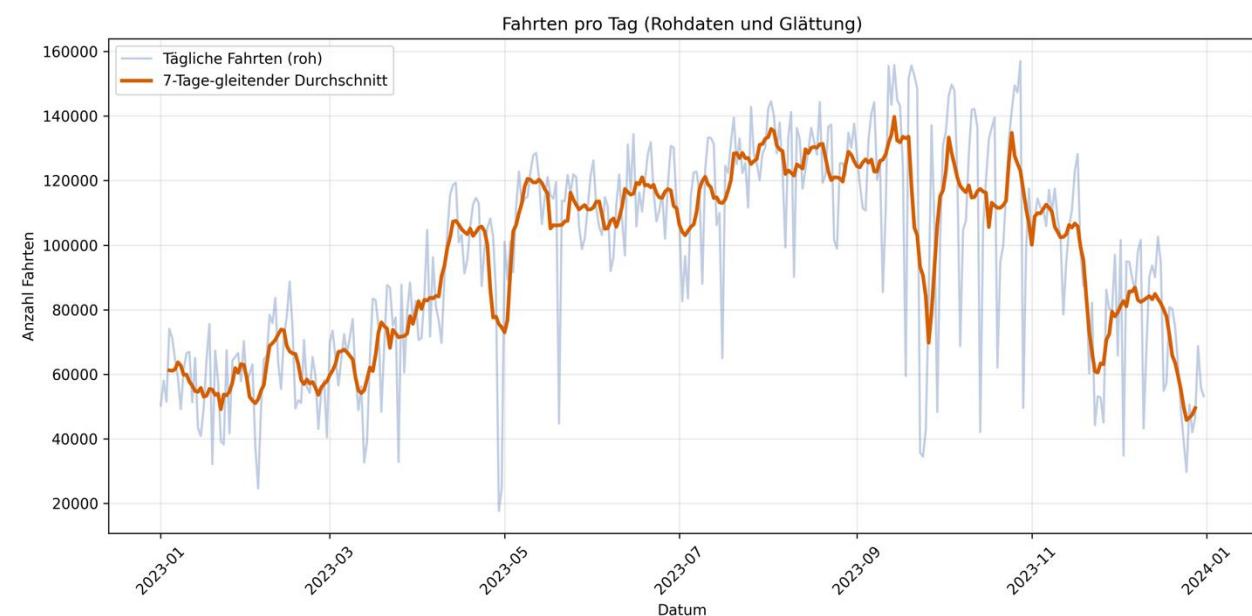
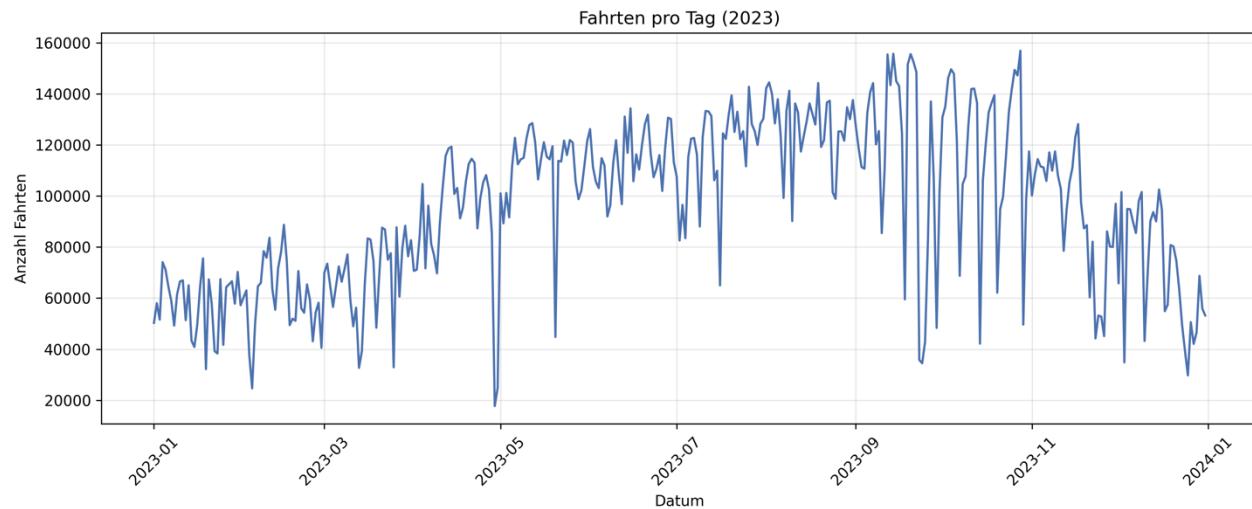
Räumliche Analyse

Phase 2B:

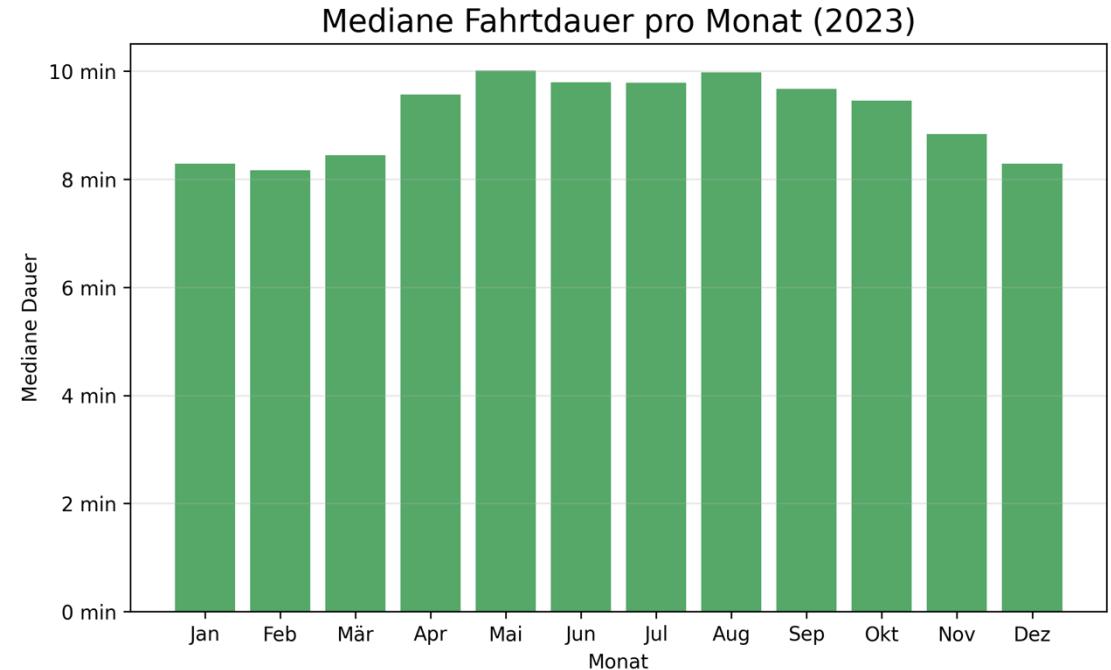
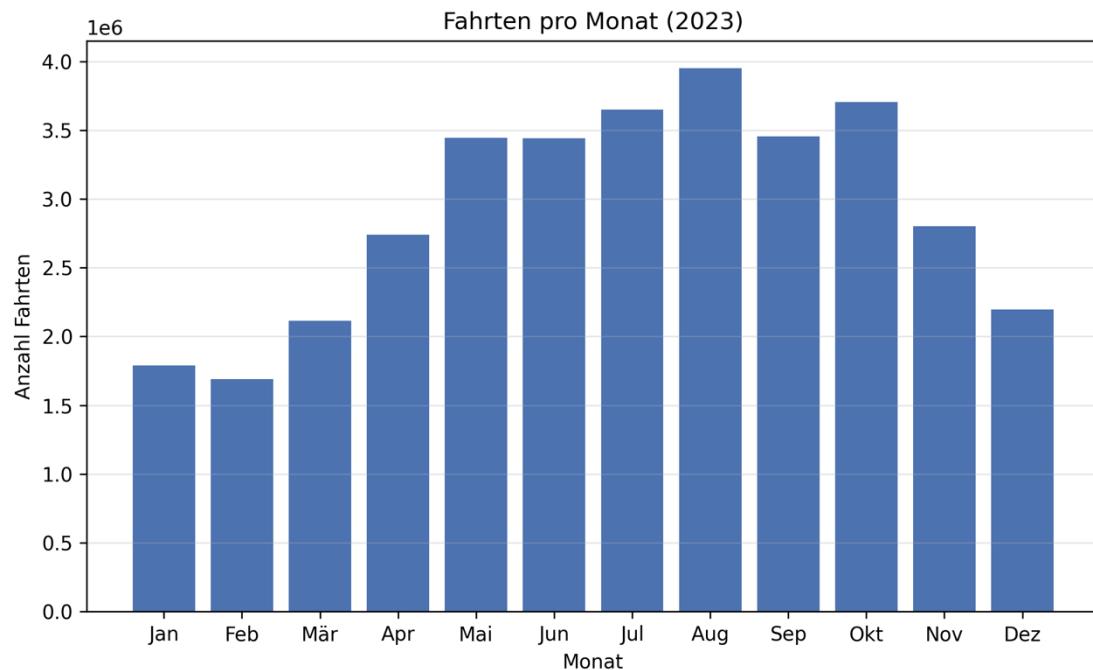
Zeitliche Analyse



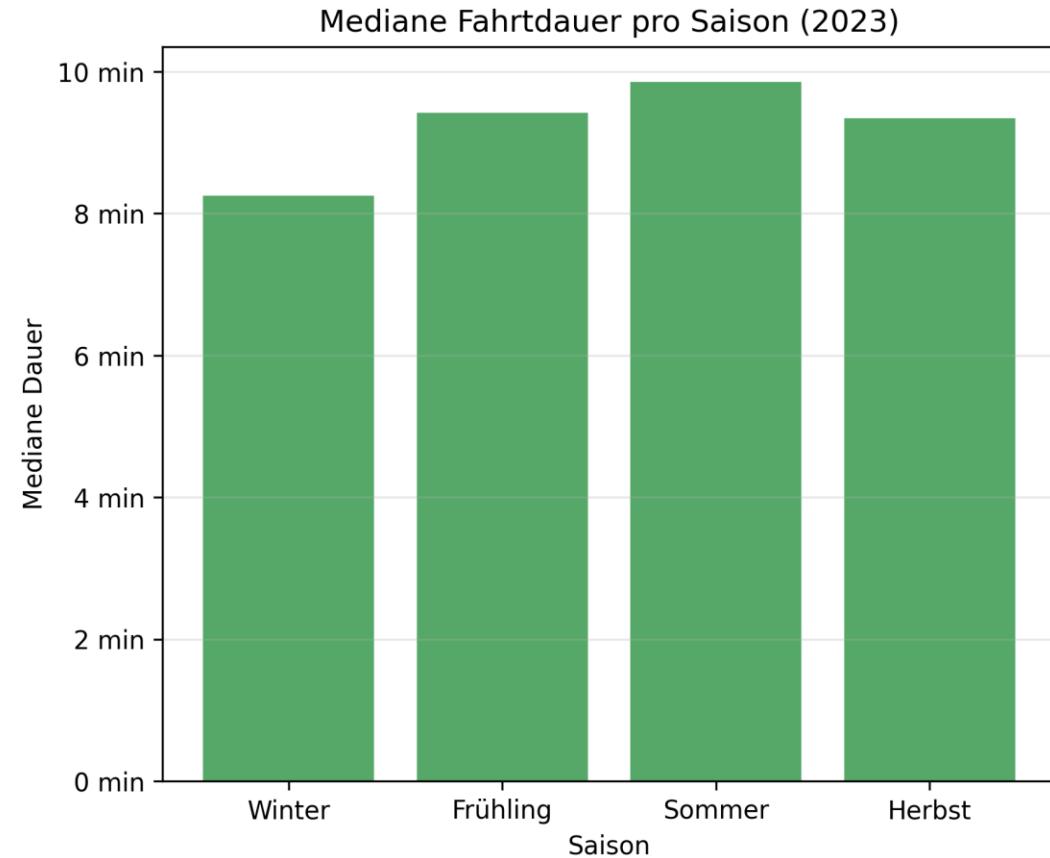
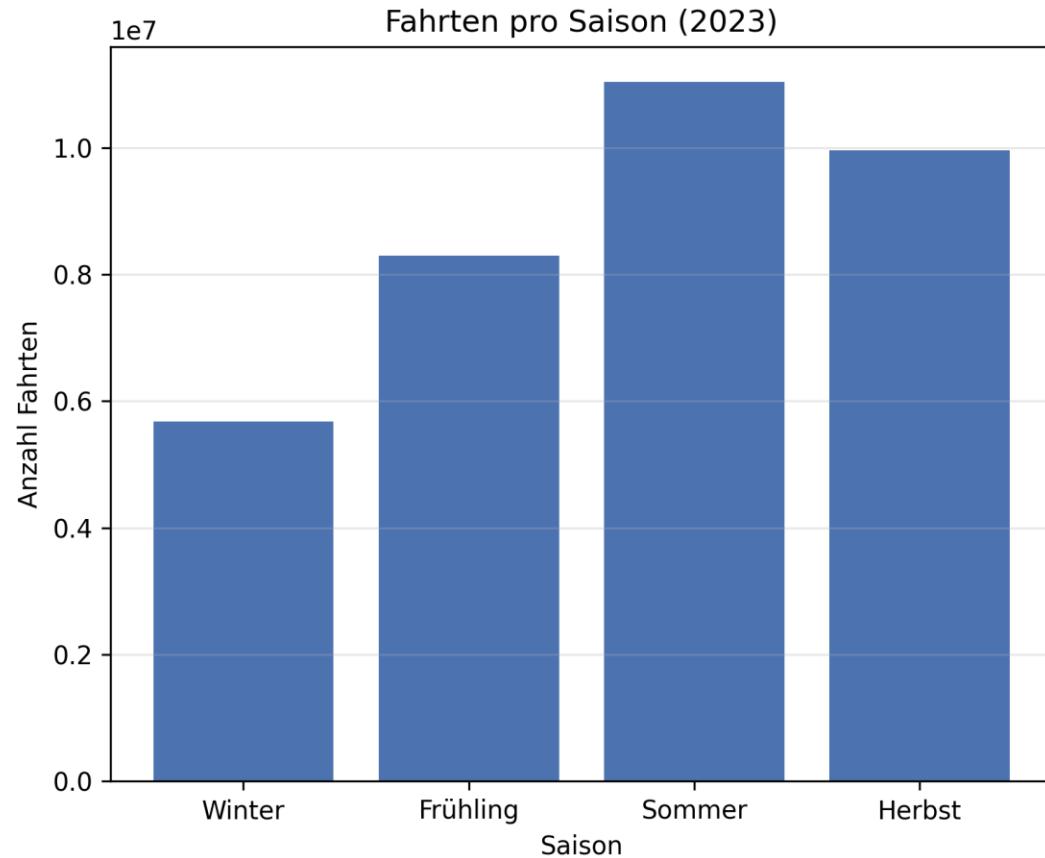
Tagesmuster



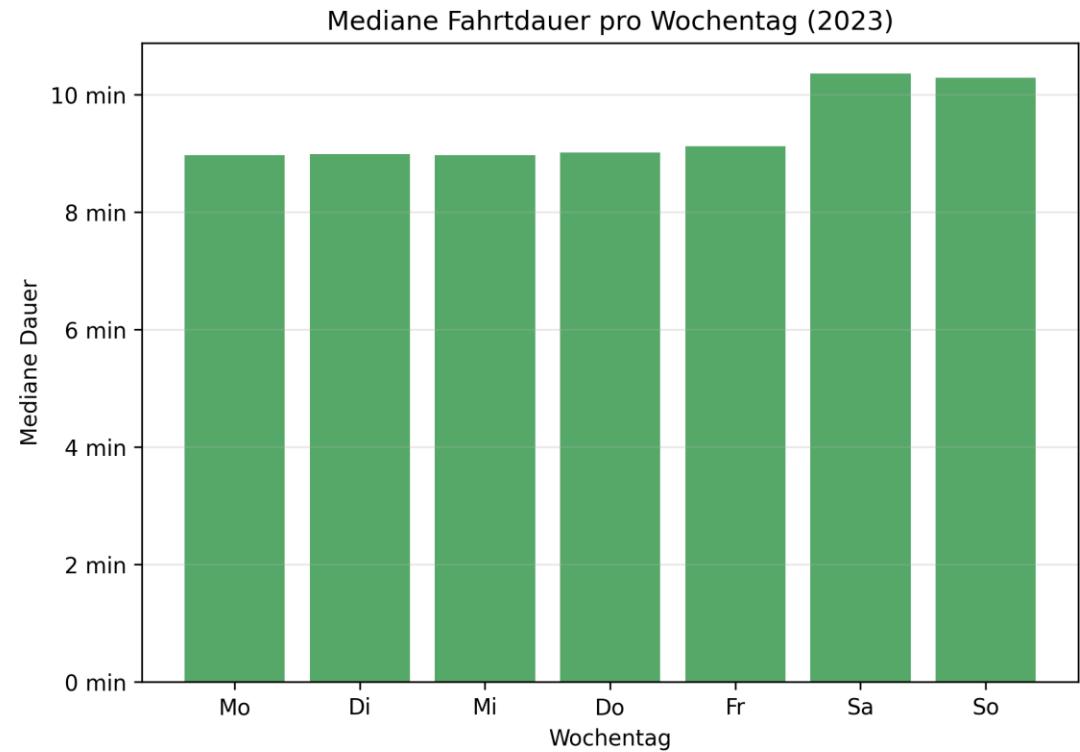
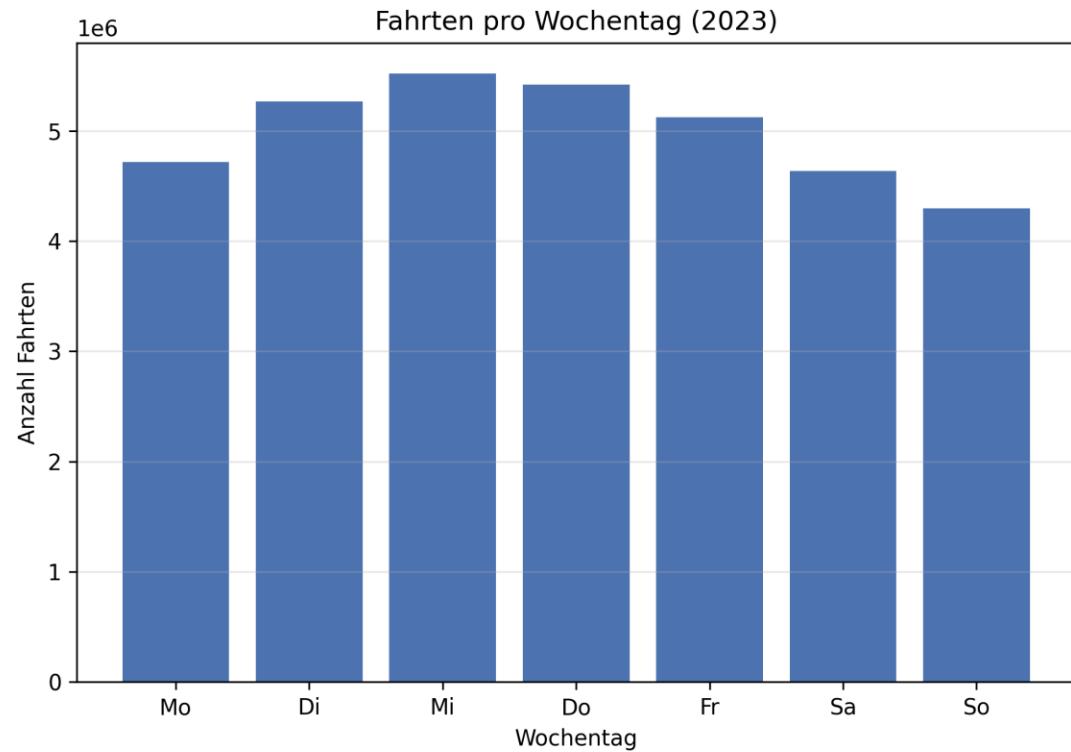
Monatsmuster



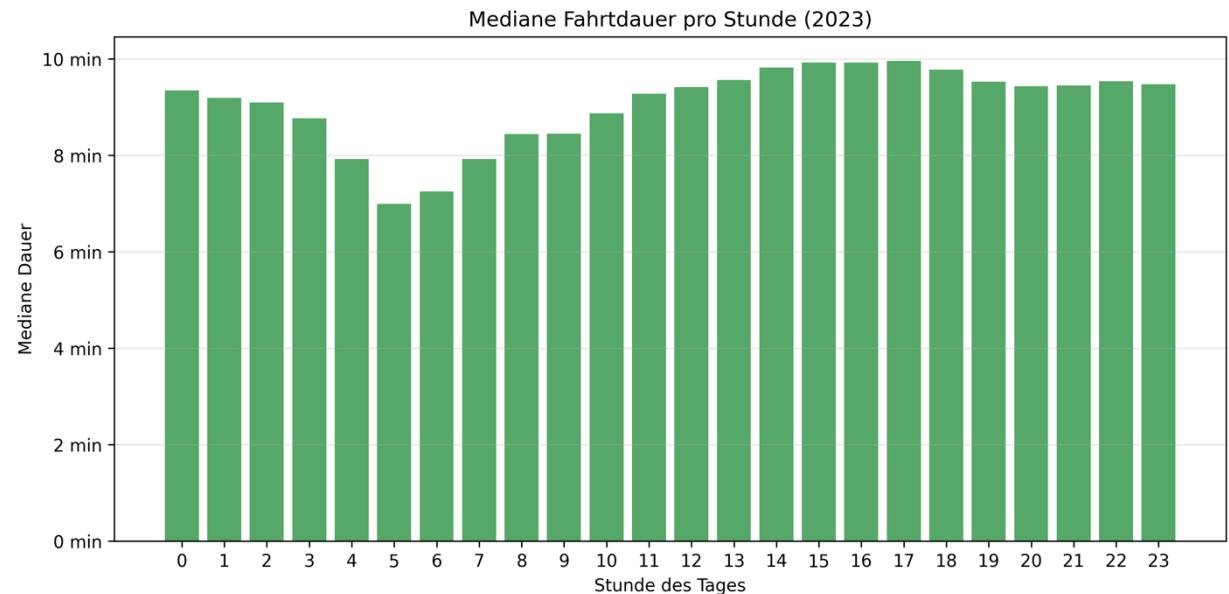
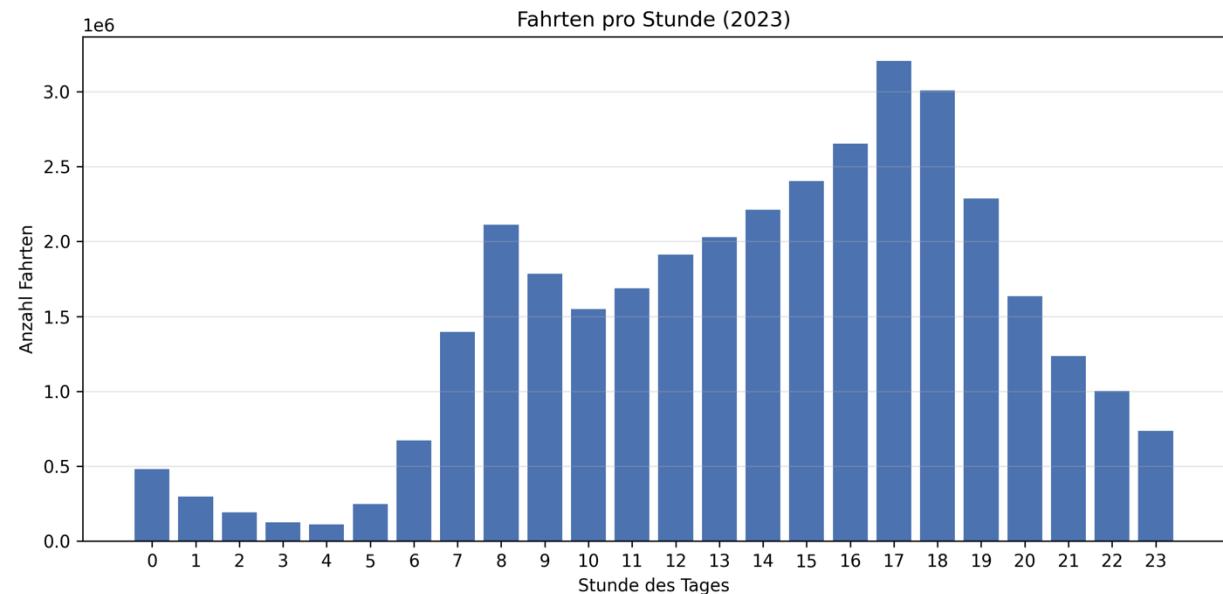
Saisonmuster



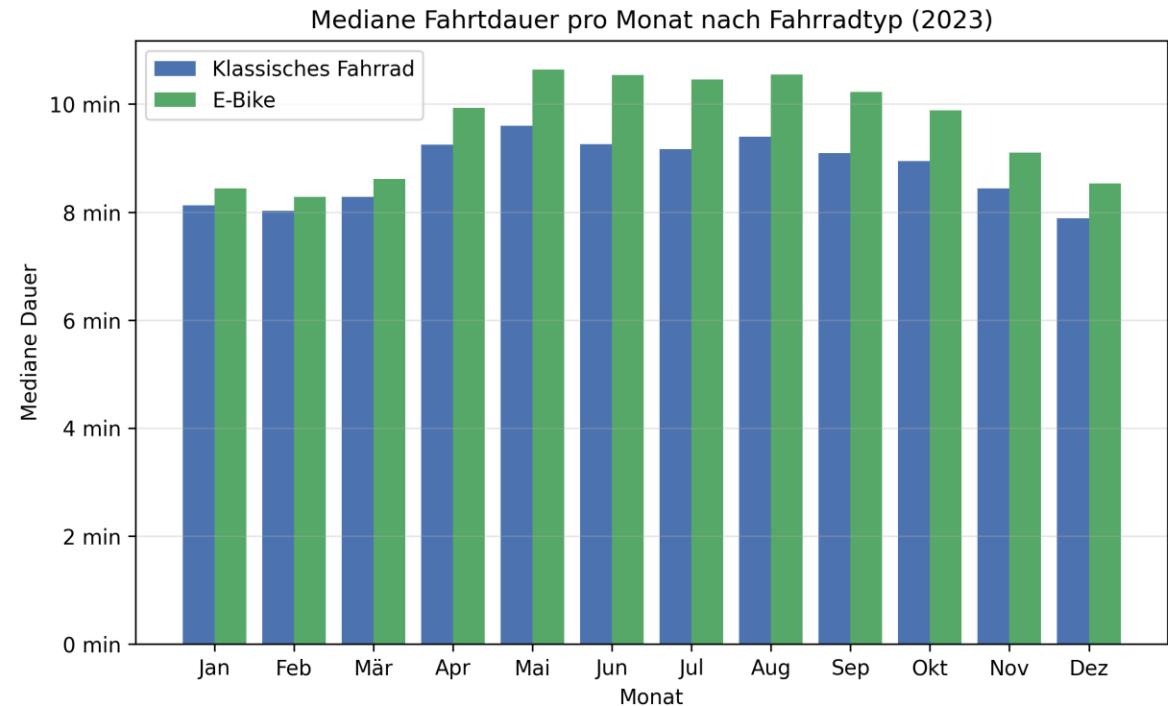
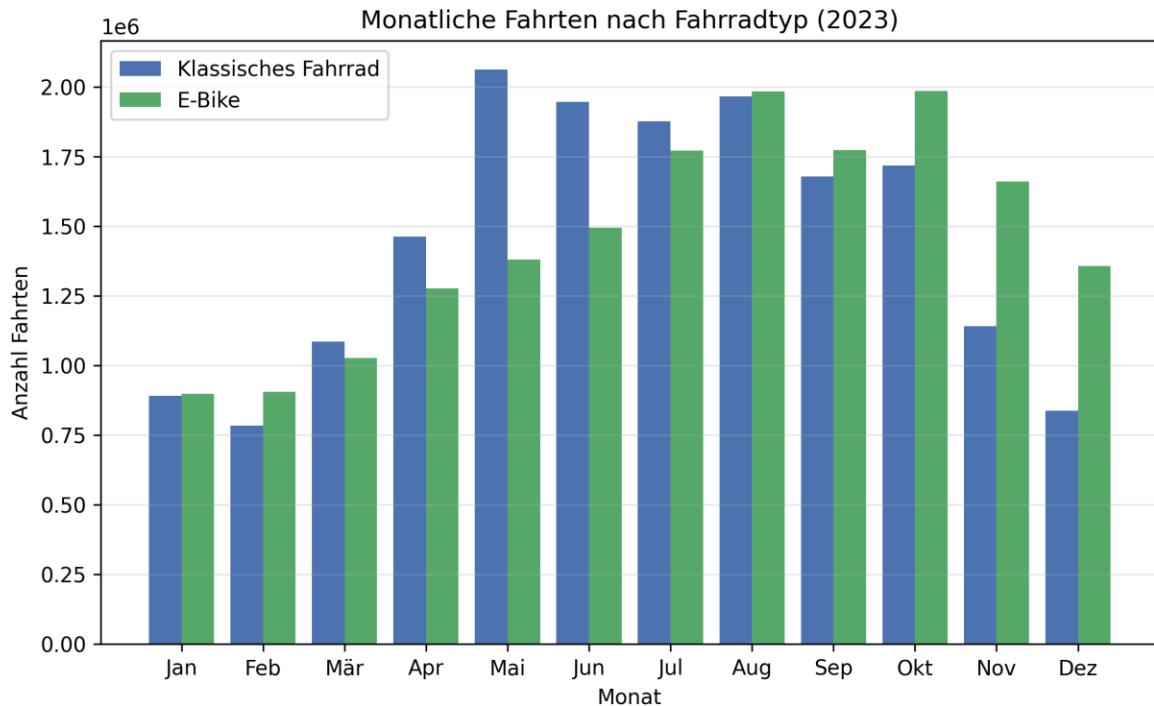
Wochentagmuster



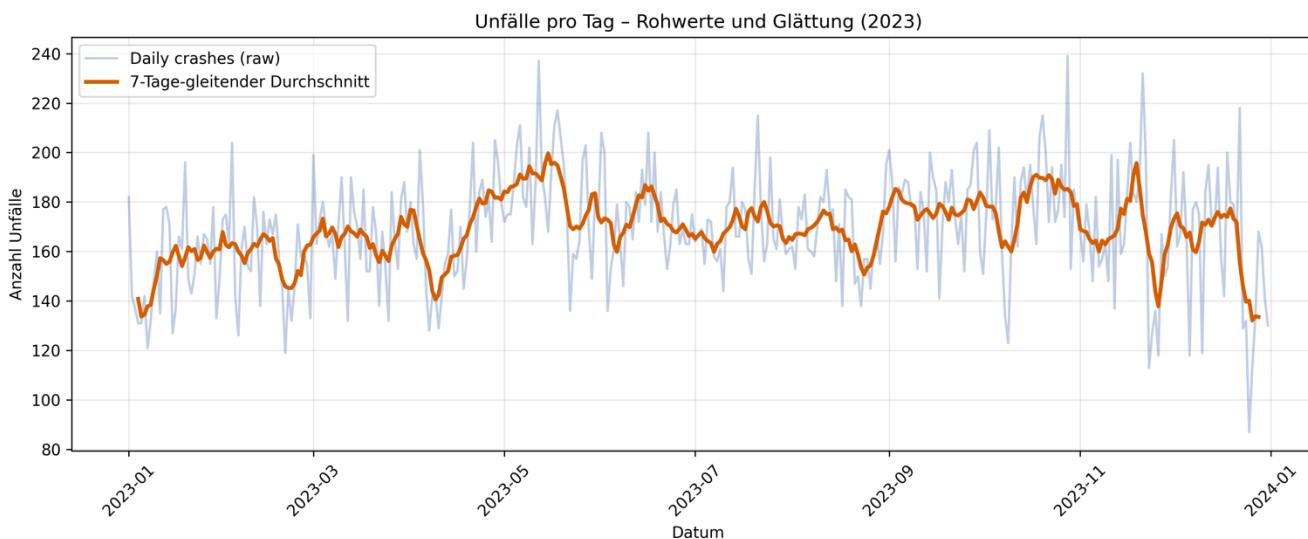
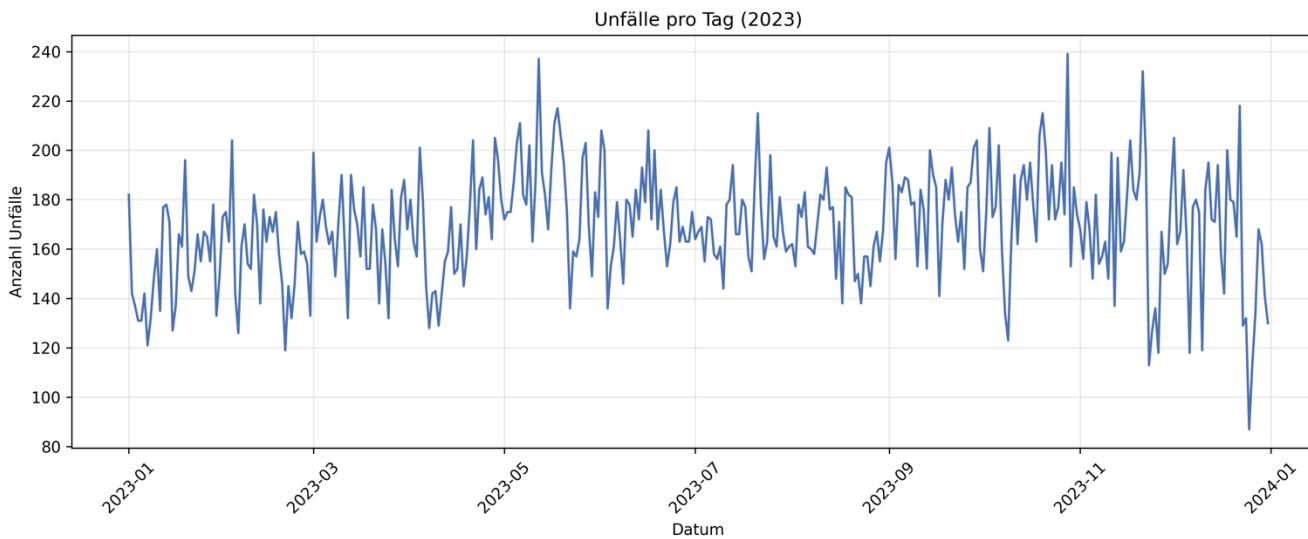
Tageszeitmuster



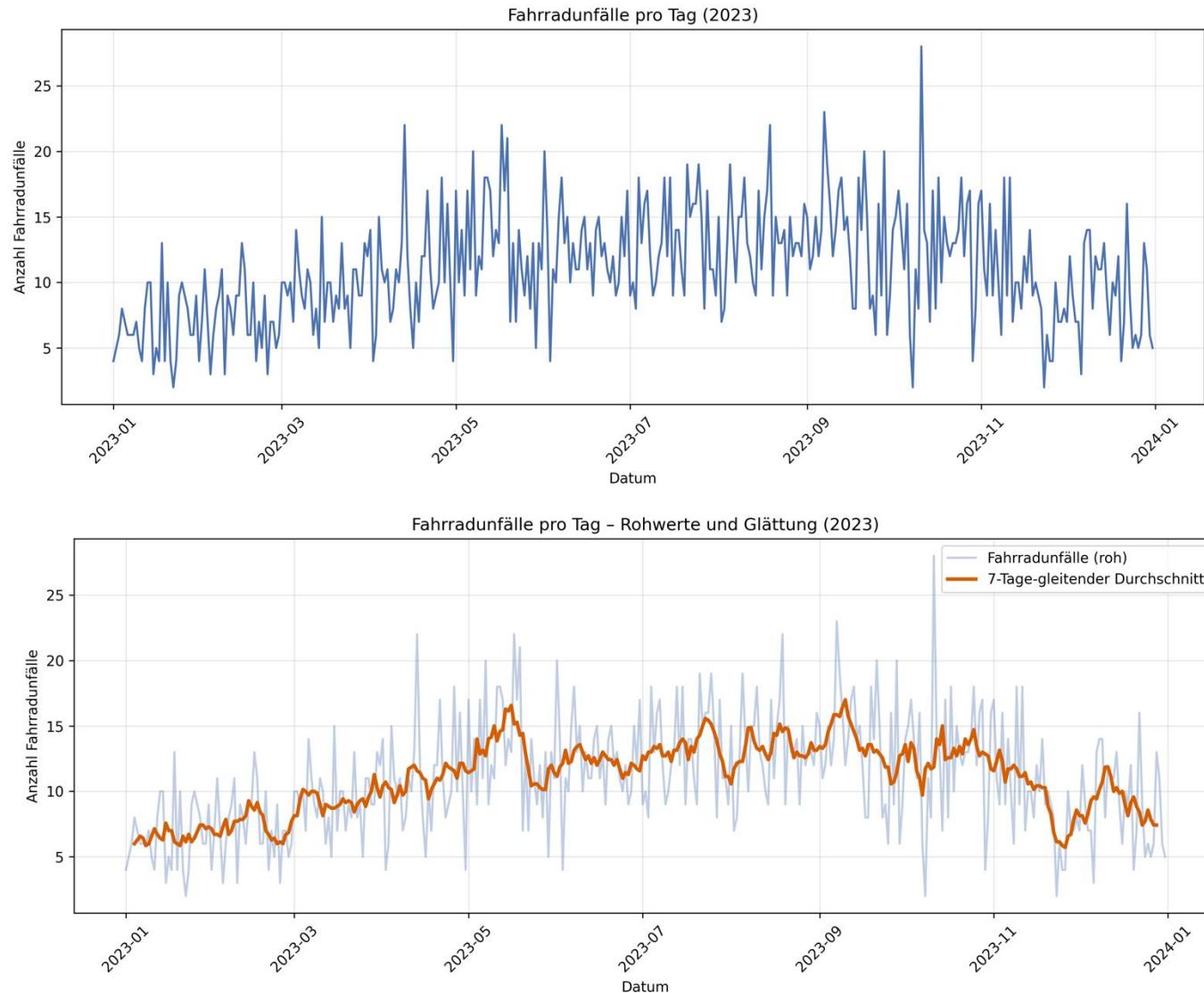
Fahrradtyp-Vergleich



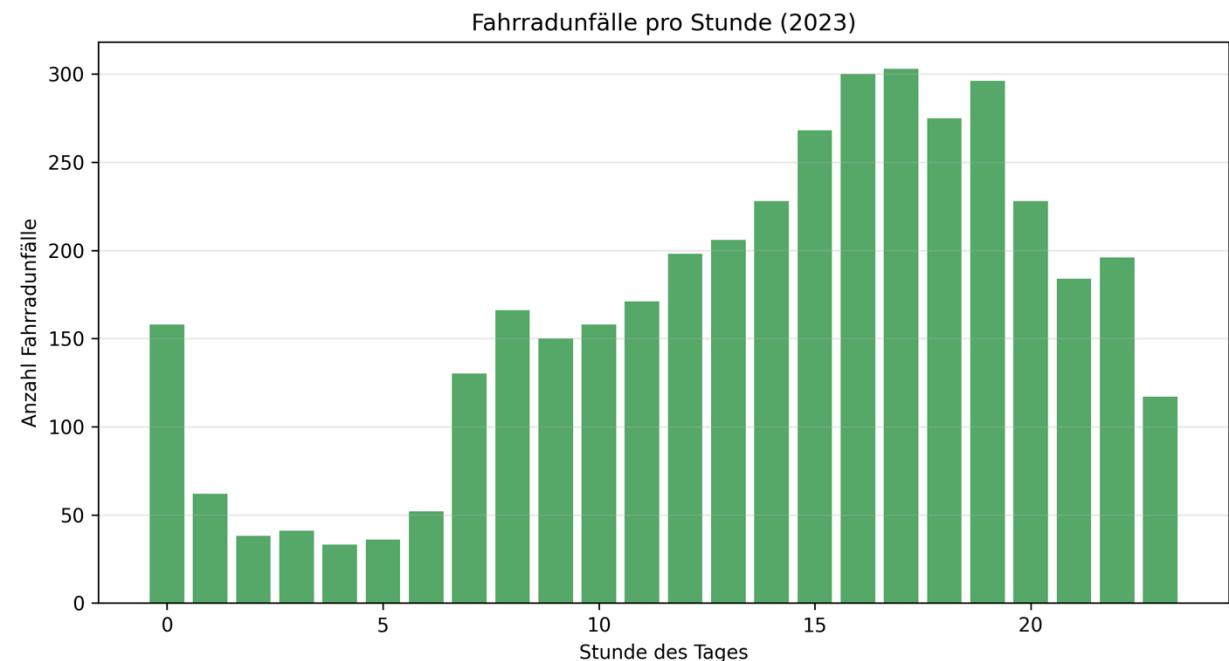
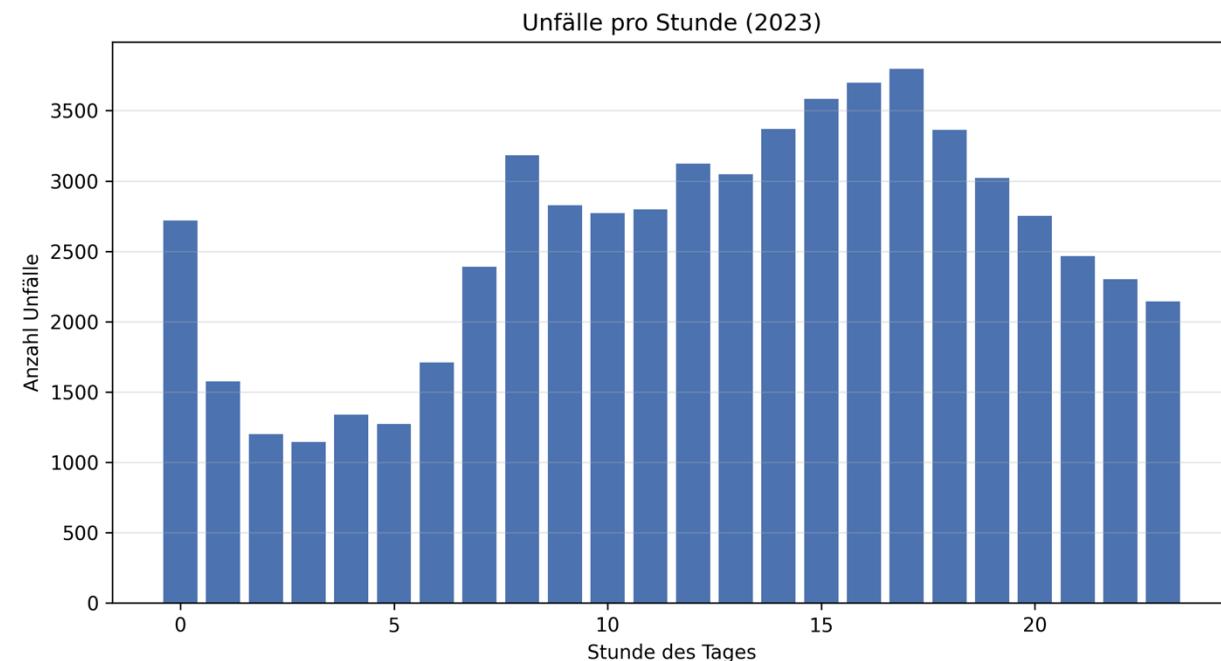
Tägliche Unfälle (Zeitreihe)



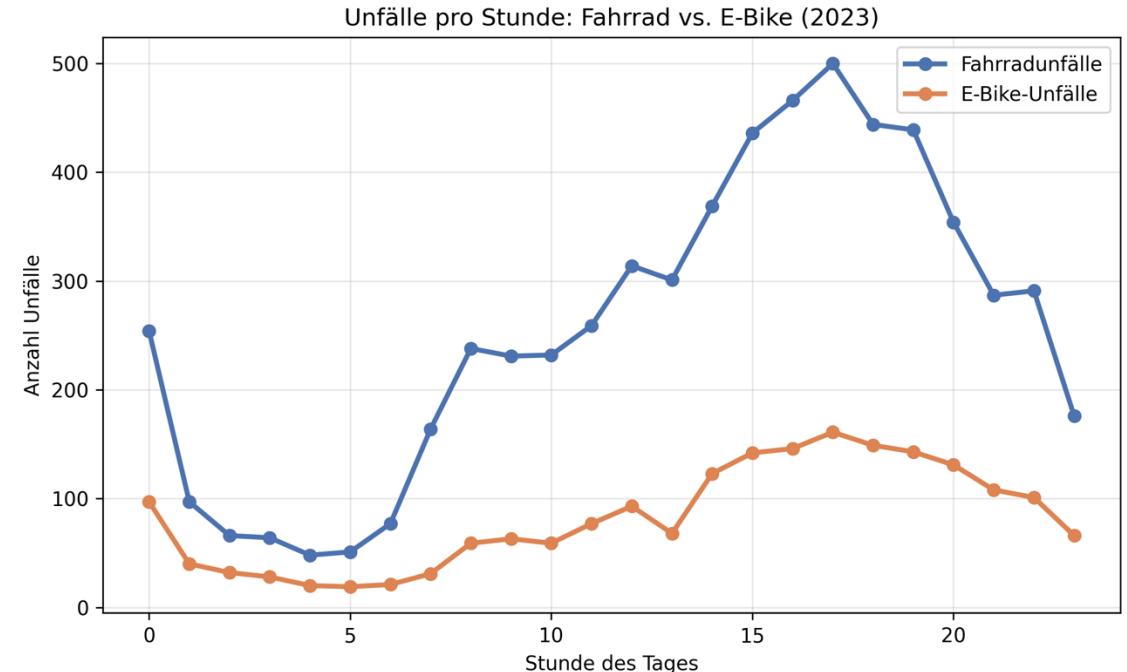
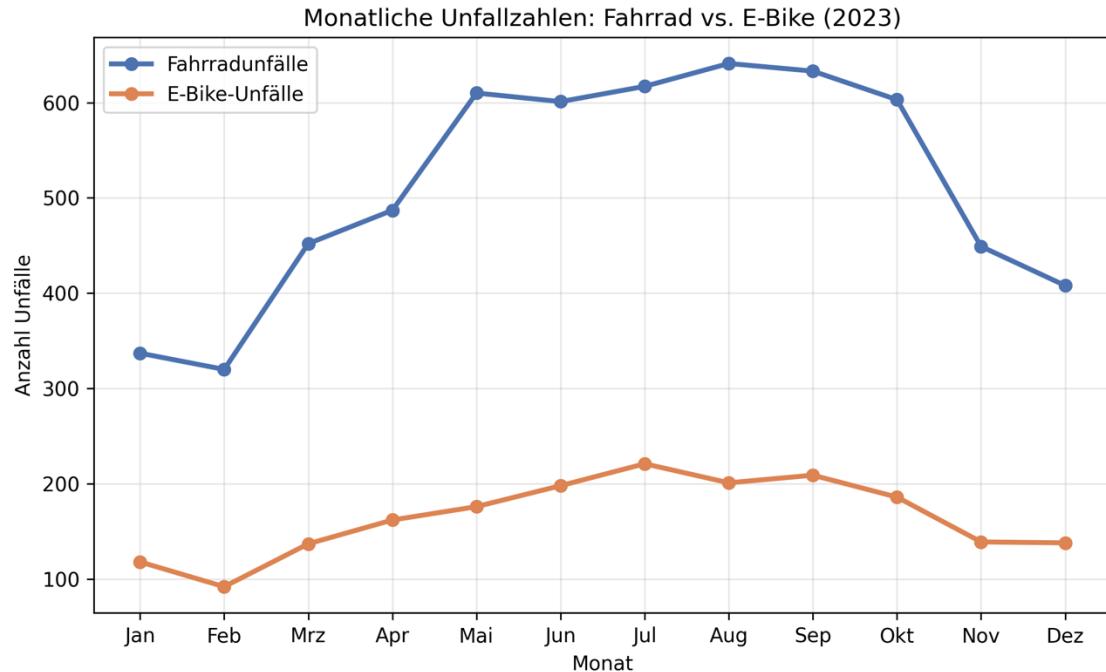
Tägliche Fahrradunfälle (Zeitreihe)



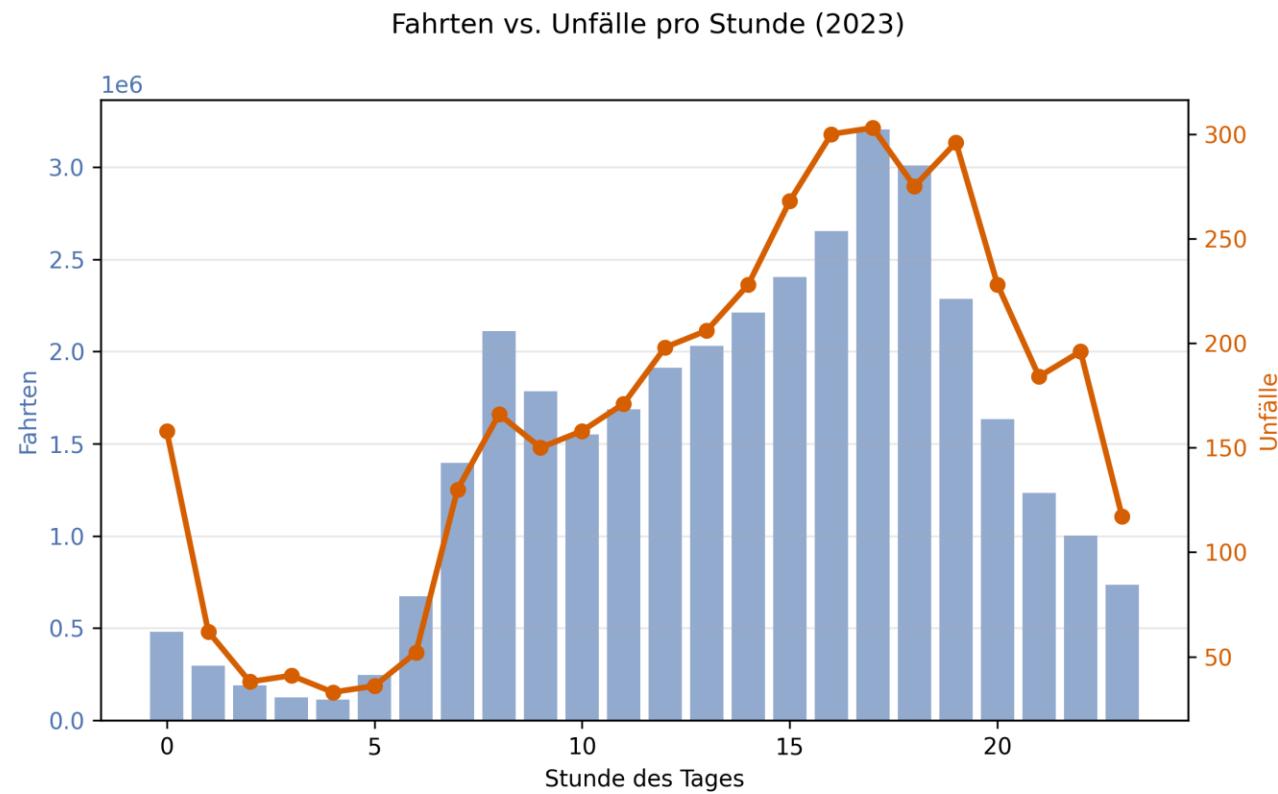
Tagesmuster



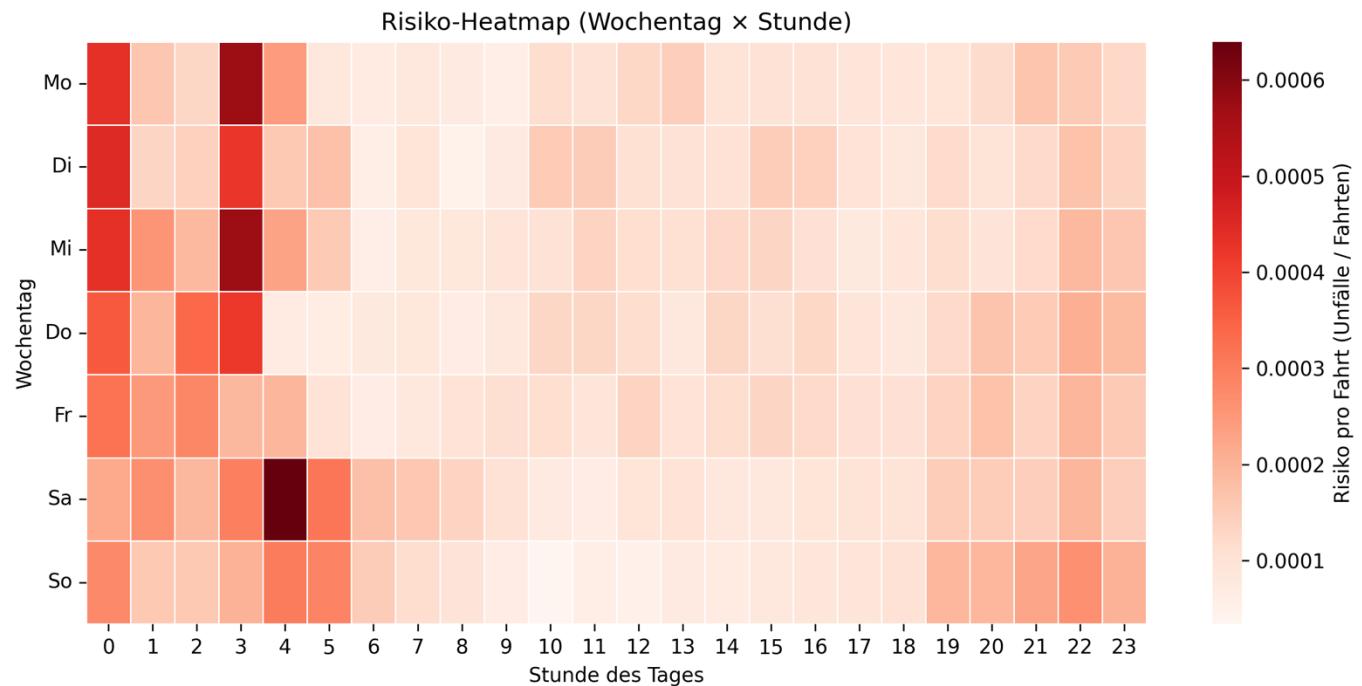
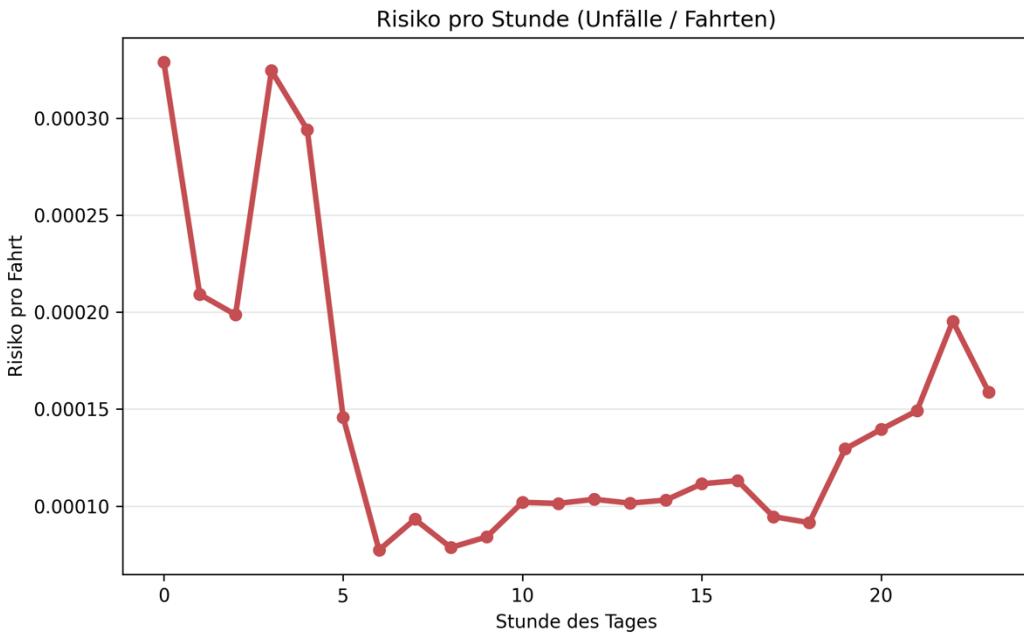
Fahrrad-Typ Vergleich



Gemeinsame zeitliche Analyse (CitiBike x NYPD)



Risikomuster



Methodisches Vorgehen

Phase 1:
Projekt-Framing

Phase 2:
Explorative Datenanalyse

Phase 3:
Hypothesentests

Phase 4:
Risikomodellierung



Warum Hypothesentests nach der EDA?

- Bis hierhin zeigt die EDA klare Muster – aber das sind **Hinweise**, keine belastbaren Risikotreiber.
- Für Micro-Insurance brauchen wir **verlässliche, stabile Zusammenhänge**, nicht visuelle Eindrücke.
- Die EDA hat Kandidaten geliefert: Ort, Zeit, Bike-Typ, ...
- Hypothesentests prüfen jetzt: **Welche dieser Muster halten einer statistischen Prüfung stand?**

Rahmen für die Hypothesentests

- Wir testen nur Muster, die in der **EDA klar sichtbar** waren.
- Wir testen nur Variablen, die **vor Fahrtbeginn bekannt** sind und in beiden Datensätzen vorkommen: Ort, Zeit, Bike-Typ.
- Wir testen jeden Faktor **direkt** auf seinen Einfluss auf die Unfallrate pro Fahrt – genau die Zielgröße, die später modelliert wird.
- Wir nutzen ein **Poisson-GLM**, weil es Crash-Counts und Exposure in einem Modell vereint und damit genau das liefert, was wir testen wollen: **Effekte auf die Unfallrate pro Fahrt**.

Hypothese 1: Safety in Numbers (Räumlicher Effekt)

- H_0 : Unfallrate ist unabhängig vom Exposure Decile.
- H_1 : Höheres Exposure-Decile → geringere Unfallrate pro Fahrt.
- Formale Modellgleichung: $\log(\mu_i) = \log(\text{exposure}_i) + \beta_0 + \beta_1 \cdot \text{exposure_decile}_i$
- $\exp(\beta_1) = 0.69$: Die Unfallrate sinkt pro Exposure-Decile um rund 30%.
- $p \approx 0$: Der Effekt ist statistisch eindeutig.
- H_0 wird abgelehnt.

=====
Kompakte Modell-Summary
=====

Anzahl Beobachtungen: 1716
Log-Likelihood: -3185.27
Overdispersion: 44.889
(≈1 gut, 1–3 normal, >4 auffällig)

feature	coef	exp(coef)	p_value	ci_2.5%	ci_97.5%
Intercept	-7.517	0.001	0	-7.611	-7.423
exposure_decile	-0.373	0.688	0	-0.388	-0.359

Methodisches Vorgehen

Phase 1:
Projekt-Framing

Phase 2:
Explorative Datenanalyse

Phase 3:
Hypothesentests

Phase 4:
Risikomodellierung

Ableitung des Modells aus den Hypothesentests

- Die als relevant eingestuften Features kommen in das Modell.
- Modelliert wird die Crashrate pro Fahrt (log-Offset für Exposure).

- Ort (Exposure-Deciles):**
Deutlich fallende Unfallraten über die Deciles → **klarer Safety-in-Numbers-Effekt.**
- Zeit (Hour/Season):**
Starke Stundeneffekte, Saison moderat → **Zeit ist ein wesentlicher Risikotreiber.**
- Bike-Typ:**
E-Bikes ~13 % der Classic-Bike-Rate → **starker Typ-Effekt.**

Kompakte Modell-Summary

Anzahl Beobachtungen: 274993
 Log-Likelihood: -13554.81
 Overdispersion: 1.598
 (≈1 gut, 1-3 normal, >4 auffällig)

feature	coef	exp(coef)	p_value	ci_2.5%	ci_97.5%
Intercept	-4.899	0.007	2.67e-84	-5.393	-4.405
C(exposure_decile)[T.1.0]	-0.726	0.484	0.00407	-1.222	-0.231
C(exposure_decile)[T.2.0]	-1.121	0.326	4.7e-06	-1.601	-0.641
C(exposure_decile)[T.3.0]	-1.402	0.246	7.4e-09	-1.877	-0.927
C(exposure_decile)[T.4.0]	-1.553	0.212	8.43e-11	-2.022	-1.084
C(exposure_decile)[T.5.0]	-2.071	0.126	5.1e-18	-2.540	-1.602
C(exposure_decile)[T.6.0]	-2.300	0.100	2.99e-22	-2.765	-1.835
C(exposure_decile)[T.7.0]	-2.762	0.063	1.77e-31	-3.226	-2.298
C(exposure_decile)[T.8.0]	-3.196	0.041	6.31e-42	-3.658	-2.735
C(exposure_decile)[T.9.0]	-3.632	0.026	2.58e-54	-4.091	-3.173
C(hour)[T.1]	-0.558	0.572	0.00453	-0.944	-0.173
C(hour)[T.2]	-0.441	0.644	0.0455	-0.873	-0.009
C(hour)[T.3]	-0.032	0.968	0.882	-0.458	0.393
C(hour)[T.4]	-0.459	0.632	0.0806	-0.974	0.056
C(hour)[T.5]	-0.927	0.396	6.2e-05	-1.381	-0.474

Vergleich: Beobachtete vs. Modellierte Crashrate

Wie würde man eigentlich die Modellgüte bewerten?

- Mit echten Out-Of-Sample Daten (z.B. 2024)

Wie wurde hier die Modellgüte bewertet?

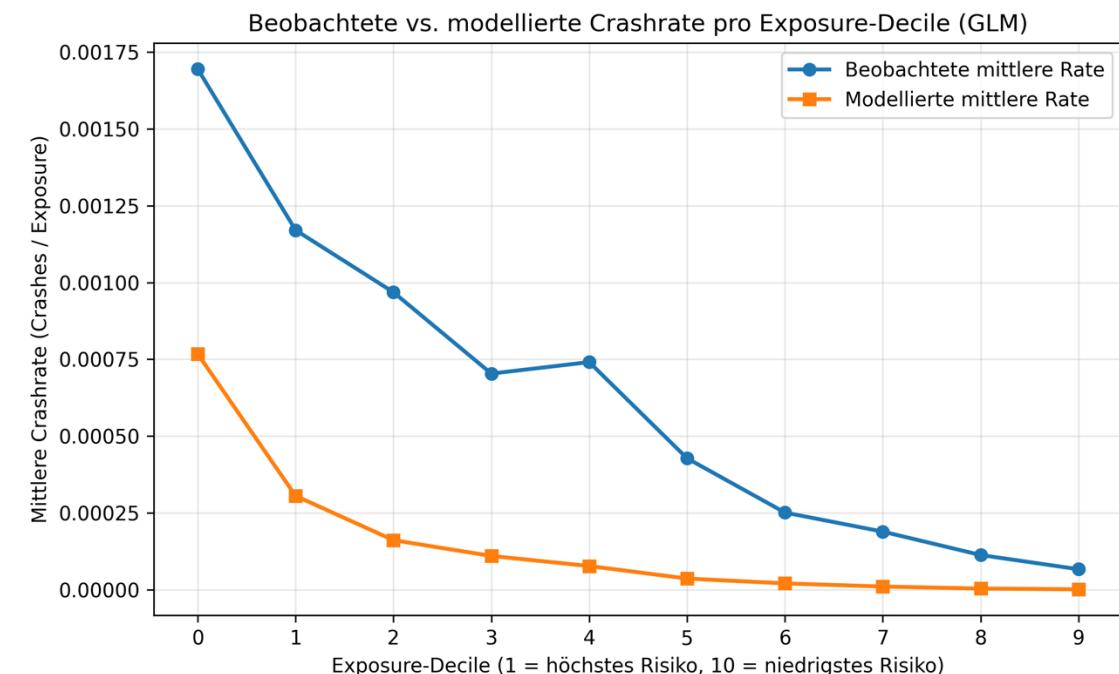
- Vergleich von beobachteter vs. vorhergesagter Crashrate pro Exposure-Decile (2023)

Was fällt auf?

- Der modellierte Trend folgt der EDA (Safety-in-Numbers-Effekt)
- Die Höhe der Crashrate wird unterschätzt

Warum ist das zunächst ausreichend?

- Die **Struktur** wird richtig getroffen
- Transparentes, interpretierbares Modell
- POC



Fazit

- Ziel war die Entwicklung eines **Analysepfades** – von den Rohdaten bis zu einem ersten Risikomodell.
- Die zentralen Muster konnten identifiziert und in ein **nachvollziehbares Modell** überführt werden.
- Der Ansatz ist **funktionsfähig**, aber bewusst einfach gehalten und klar **ausbaufähig**.
- Wenn das Produkt für CitiBike interessant ist, ließe sich darauf eine **nächste Iteration** aufbauen.

Vielen Dank für eure Aufmerksamkeit ☺